SISTEMA PARA EL ANÁLISIS DE DATOS DE CONSUMO DE ENERGÍA ELÉCTRICA BASADO EN TÉCNICAS DE INTELIGENCIA DE NEGOCIOS Y MINERÍA DE DATOS

LEIDY DANIELA MINA PALACIOS

FACULTAD DE INGENIERÍA
ESCUELA DE INGENIERÍA DE SISTEMAS Y COMPUTACIÓN
PROGRAMA ACADÉMICO DE INGENIERÍA DE SISTEMAS
SANTIAGO DE CALI, ABRIL DE 2019

SISTEMA PARA EL ANÁLISIS DE DATOS DE CONSUMO DE ENERGÍA ELÉCTRICA BASADO EN TÉCNICAS DE INTELIGENCIA DE NEGOCIOS Y MINERÍA DE DATOS

Modalidad Desarrollo

Leidy Daniela Mina Palacios
201510360
leidy.daniela.mina@correounivalle.edu.co

Director
Oswaldo Solarte Pabón
Magíster en Ingeniería de Sistemas
oswaldo.solarte@correounivalle.edu.co

Facultad de Ingeniería
Escuela de Ingeniería de Sistemas y Computación
Programa Académico de Ingeniería de Sistemas
Santiago de Cali, Abril de 2019

Agradecimientos

A Dios por darme la fortaleza física, mental y espiritual para cumplir con este meta. Porque solo por su misericordia y bondad pude tener vida y salud para culminar con este reto de tanta importancia en mi vida. A mis padres Orlando Mina y Maria Eugenia Palacios por apoyarme incondicionalmente en cada aspecto de mi vida. Por poner siempre en primer lugar mi bienestar.A mi director de tesis Oswaldo Solarte por el apoyo y orientación brindada en este proceso. A Jenniffer Guerrero por su constante disposición a ayudarme y orientarme a lo largo de este proceso. A Jean Micolta por su apoyo incondicional. A todas las personas que han contribuido de diversas formas a mi formación académica y personal.

Nota de aceptación

_____

_____

_____

_____

_____

_____

Firma del Presidente del Jurado

_____

Firma del Jurado

_____

Firma del Jurado

Santiago de Cali, Abril de 2019

Índice

Índice de tablas

Índice de figuras

1. Resumen

Tradicionalmente, la medición del consumo de energía eléctrica residencial se ha realizado por medio de medidores análogos para determinar posteriormente el costo de facturación. Sin embargo actualmente está en ejecución la implementación de actualizaciones de este proceso por medio de la implementación de la Infraestructura de Medición Avanzada (AMI) como una respuesta a las necesidades de optimización de recursos de prestación de servicio y aprovechamiento de los avances en tecnologías de información.

AMI permite tener acceso a datos relacionados con el consumo eléctrico de los usuarios con tiempos de latencia relativamente bajos con respecto al momento de su generación; debido a esto se generan grandes cantidades de datos que requieren de herramientas para ayudar a analizar estos datos. Los datos se encuentran generalmente en archivos de texto, específicamente en formatos .dat o .csv.

En este proyecto se propone la creación de un sistema que implemente técnicas de inteligencia de negocios y minería de datos con el objetivo de procesar los datos generados por sistemas AMI y que permita la creación de ventajas competitivas por medio del análisis de estos datos. Se busca crear una herramienta que facilite la integración de diferentes archivos donde se encuentran los datos generados por sistemas AMI, la creación de módulos de ETL (Extract, Transform and Load) y mecanismos de visualización de reportes.

El resto del documento está organizado de la siguiente forma: En la Sección 2 se presenta el planteamiento y justificación del problema, en la Sección 3 se muestran los objetivos del proyecto, en la sección 4 se describe el marco referencial y en la Sección 5 la metodología y el presupuesto.

Parte I
Descripción general

En esta parte del documento se describe de forma general el problema y la metodología propuesta para tratarlo.Primero, se describe el problema a tratar. Después, se detalla la importancia de implementar un solución en el área de sistemas de información para dicho problema y objetivos a alcanzar medio de la implementación de este trabajo. Finalmente, se describe el alcance del trabajo.

2. Planteamiento y Formulación del Problema

Internet de las cosas (Internet-of-Things IoT) ha dejado ser solo un tema de investigaciones para convertirse en una realidad que ha permitido la creación de diversos ambientes inteligentes con un abanico de posibilidades tanto para diseñadores como para usuarios y necesita de la intervención de técnicas y herramientas de las TIC para su ejecución óptima. En el marco de este desarrollo se propone la Infraestructura de Medición Avanzada que permite la comunicación bidireccional entre los usuarios y los operadores de redes eléctricas. Esta infraestructura integra hardware, software y arquitecturas y redes de comunicaciones, que permiten la operación de la infraestructura y gestionar(consultar, almacenar y compartir) datos del sistema de distribución de energía eléctrica. " Más allá de las lecturas automáticas de contadores para fines de facturación, los contadores inteligentes sirven como pasarelas de información para las instalaciones del cliente. Proporcionan lecturas horarias de intervalos de 15 minutos o más frecuentes para respaldar programas de precios dinámicos, mejorar la administración de ingresos, comunicaciones proactivas con los clientes para revelar oportunidades de eficiencia energética y conservación, así como alertas de consumo y facturación. También pueden incluir funciones de control como la desconexión / reconexión remota para agilizar las operaciones del cliente y la capacidad de limitación de demanda o prepago para ayudar a los clientes a administrar sus presupuestos y consumo de energía"[9].

En colombia, por medio del decreto 348 de 2017 se reglamentó "establecer e implementar lineamientos de política energética en materia de sistemas de medición avanzada, así como la gradualidad con la que se deberá poner en funcionamiento, con el fin de promover la gestión eficiente de energía, y promover la incorporación de tecnologías de redes inteligentes" [10]; Con la llegada de los sistemas AMI la cantidad de datos intercambiados en la red aumenta de manera significativa gracias a la continua comunicación en la red bidireccional entre usuarios y proveedores del servicio y aunque es relativamente fácil la obtención de dichos datos por parte de las compañías energéticas, se complican los procesos de gestión y obtención de conocimiento a partir de ellos. Por esto se hace necesaria la implementación de tecnologías de la información que permitan simplificar los procesos de gestión de datos generados por sistemas AMI. Entre las áreas

que proveen herramientas de gestión de información útiles, se encuentran la inteligencia de negocios y la minería de datos.

3. Justificación del Problema

La implementación de técnicas de Inteligencia de Negocios(BI) y Minería de Datos(DM) como solución al problema planteado mejoraría notablemente los procesos de medición de consumo de electricidad, relación con el cliente, manejo de información y la toma de decisiones. Además de permitir la implementación de medidas de control de consumo basadas en la información provista por el sistema.

Justificación Académica

La realización de este proyecto permite la práctica y fortalecimiento de habilidades en diversas áreas de estudio de ingeniería de sistemas; principalmente de desarrollo de software y el descubrimiento de conocimiento en bases de datos. Además de propiciar una actualización en tecnologías y procesos que fortalecen el perfil profesional, generando un amplio rango de ventajas competitivas frente a profesionales de la misma área.

Justificación Económica

Con el crecimiento del uso de tecnologías en diversas áreas, se han abierto amplios campos de investigación y aplicación de tecnologías de la información; entre estas áreas se encuentra la industria eléctrica que ha experimentado un gran crecimiento en las últimas décadas y ha generado grandes oportunidades no solo para profesionales en electricidad, sino también para profesionales en gestión de información. Esta industria propone nuevos retos que pueden ser afrontados de manera eficiente y eficaz por medio de la implementación de técnicas de diversas ramas de estudio de la ingeniería de sistemas.Por todo esto, la implementación de esta propuesta de trabajo de grado es una oportunidad potencial de incursión en el mundo laboral, además de ser un gran aporte a la industria eléctrica del país.

4. Objetivos

Objetivo General

Desarrollar una herramienta para analizar datos del consumo de electricidad en usuarios residenciales usando técnicas de inteligencia de negocios y minería de datos.

Objetivos Específicos

1. Explorar herramientas para almacenamiento y carga de grandes cantidades de datos como estructuras de bodegas de datos.

2. Diseñar una bodega de datos para almacenamiento de información de consumo de energía.

3. Desarrollar un módulo de extracción transformación y carga de datos (ETL).

4. Desarrollar un módulo de visualización y reportes para la información obtenida de la implementación de técnicas de minería de datos e inteligencia de negocios sobre datos de consumo de energía eléctrica.

5. Evaluar los módulos desarrollados mediante el diseño e implementación de pruebas de software.

Resultados Esperados

Objetivo	**Producto(s) Esperados**
Objetivo 1	Reporte comparativo de las diferentes herramientas exploradas. En la Sección 9 se detalla la investigación de herramientas de almacenamiento de grandes volúmenes de datos.
Objetivo 2	Documentación del diseño de la bodega de datos. En Sección 10 se detalla las fases y componentes del diseño de la bodega de datos.
Objetivo 3	Código fuente de módulo ETL. En la Sección 11 se describe y visualiza por medio de diagramas las fases del proceso de Extracción, Transformación y Carga de los datos.
Objetivo 4	Código fuente de módulo de visualización y reportes. En la Sección 13 se realiza el análisis de requerimientos para la implementación del modulo, el proceso de desarrollo y las características del producto final.
Objetivo 5	Documentación de diseño de pruebas y reporte de resultados. En la Sección 14 se describen los procesos de pruebas a realizar y sus respectivos resultados.

Tabla 1: Resultados esperados por objetivo especifico

5. Alcances de la Propuesta

Se propone la realización de una herramienta web que permita el análisis de datos recopilados en los procesos de medición de consumo eléctrico residencial mediante la implementación de técnicas de inteligencia de negocios y minería de datos, con el objetivo de obtener información de utilidad relacionada con el comportamiento y características de consumo de usuarios del servicio eléctrico. Esta herramienta deberá permitir la visualización de la información generada por medio de reportes escritos y/o gráficos de fácil análisis e interpretación. En el proceso de exploración de herramientas para gestión de datos orientadas a bodegas de datos, solo se tendrán en cuenta herramientas de software libre.

Parte II
Marco Referencial

En esta parte del documento se detalla la investigación conceptual que enmarca la propuesta de solución. Primero se presentan las definiciones cortas de términos usados en el presente trabajo. Después, se analizan algunas investigaciones y trabajos experimentales en el campo de análisis de consumos de energía eléctrica por medio de tecnologías de inteligencia de negocios y minería de datos. Finalmente, se detalla la investigación conceptual de las principales tecnologías influyentes en el desarrollo del trabajo.

6. Glosario

Ambiente Inteligente

Mundo físico que está ricamente entretejido e invisible con sensores, actuadores, visualizadores y elementos computacionales, integrados a la perfección en los objetos cotidianos de nuestras vidas[11].

CRISP-DM (Cross Industry Standard Process for Data Mining)

Metodología de gestión de proyectos enfocada en la implementación de proyectos de minería de datos.

Data Warehouse

Un almacén de datos o data warehouse es "un repositorio de información coleccionada de varias fuentes, almacenada bajo un esquema unificado que normalmente reside en un único emplazamiento" [12]. El data warehouse contiene la información de varios procesos de negocio de una compañía.

Datamart

Un datamart es una versión simplificada del data warehouse y suele contener información de un solo proceso de negocio de una compañía.

Internet-of-Things(IoT)

Interconexión digital de objetos cotidianos por medio de internet.

Medidores Avanzados

Dispositivos que miden y registran datos de uso de energía eléctrica en intervalos de una hora como mínimo, con capacidad de almacenar y transmitir datos. Estos datos

se proporcionan, por lo menos, con frecuencia diaria a los operadores de red y a los usuarios. La información registrada se puede utilizar, entre otros fines, para la gestión comercial, la planeación y operación del sistema, y la gestión de pérdidas. Los medidores avanzados posibilitan la comunicación bidireccional entre el usuario y el operador de red.

Red Inteligente

Sistema de aplicaciones de información y comunicaciones integradas con la generación, transmisión, distribución, y las tecnologías de uso final de energía eléctrica. Esta red permite la participación activa de los usuarios, la integración de las opciones de generación y almacenamiento locales, la incorporación de energías renovables, la optimización y operación más eficiente del sistema de potencia, la anticipación y respuesta ante perturbaciones en el sistema, y la operación flexible contra apagones o desastres naturales, entre otros.

TIC

Tecnologías de la información y comunicación.

7. Estado del Arte

Con la creación de los diversos retos en obtención de conocimiento planteados debido a la implementación de tecnologías AMI en el sector eléctrico se han realizado diferentes investigaciones con el objetivo de maximizar los beneficios de esta tecnología no solo a nivel de procesos operativos como apoyo a la gestión y monitoreo , sino también a nivel ejecutivo como apoyo a la predicción y toma de decisiones empresariales; por medio de la aplicación de técnicas de inteligencia de negocios y minería de datos.

La minería de datos se define como el proceso no trivial de identificar patrones de datos válidos, novedosos, potencialmente útiles y, en última instancia, comprensibles[13]. A esta noción de descubrimiento se le han dado varios nombres, incluida la extracción de conocimiento, la exploración de bases de datos, el procesamiento de patrones de datos, la minería de datos, la recolección de información, el software y el dragado de datos. También se lo conoce como descubrimiento de conocimiento en bases de datos donde los algoritmos de descubrimiento se enfocan en datos almacenados en bases de datos, de modo que pueden explotar las oportunidades inherentes presentadas por la organización de datos en bases de datos. La implementación y el uso efectivos de las herramientas aún requieren una gran experiencia en la extracción, manipulación y análisis de datos de grandes almacenes de datos. Sin embargo, estas herramientas actualmente están produciendo resultados que proporcionan importantes ventajas estratégicas y tácticas para las empresas en entornos altamente competitivos. A medida que estas herramientas continúan madurando, la interfaz de usuario mejora; una mayor población de analistas familiarizados con estas herramientas llevará la extracción de datos a la corriente principal de la tecnología de la información y las aplicaciones de ingeniería, incluidos los sistemas de energía[14].

El trabajo realizado por Fengyu Wang, Yanliu Cheng, Houlei Lv y Yingxin Xie presenta el sistema de inteligencia de negocio de consumo de energía inteligente, el cual utiliza tecnología de computación en la nube para implementar el almacenamiento distribuido y la computación paralela de datos masivos del nivel de TB. Se basa en sistemas de archivos distribuidos y en un marco informático paralelo, y en conectar fuentes de datos heterogéneas al almacén de datos en la nube mediante la tecnología ETL paralela. La consulta ad hoc, el análisis OLAP(Procesamiento Analítico En Linea) multidimensional, la extracción de datos y la función de aprendizaje automático proporcionados por el sistema de inteligencia de negocio de consumo de energía inteligente ha implementado aplicaciones multinivel de monitoreo de calidad de energía en tiempo real, análisis multidimensional del indicador de consumo de energía, análisis de comportamientos de consumo de energía de los usuarios, e implementar inteligencia de negocios en el consumo de energía. El almacenamiento de datos distribuidos resuelve el problema de almacenamiento de datos masivos de consumo de energía inteligente. Y la base del marco de cómputo paralelo en MapReduce también proporciona la solución para el procesamiento de análisis y la minería de datos en profundidad de los datos masivos[15].

El trabajo realizado por N. Jacome-Grajales, G. Escobedo-Briones, J. Roblero y G. Arroyo-Figueroa describe el diseño y desarrollo de herramientas de inteligencia de negocios aplicadas a la seguridad industrial. El objetivo es tener información disponible y oportuna para tomar mejores decisiones y reducir el número de accidentes e incidentes industriales, estas herramientas pueden ser usadas en el futuro para procesos relacionados con las prestación de servicios de energía eléctrica. Este trabajo concluye que las herramientas de inteligencia de negocios desarrolladas han tenido buenos resultados, la información que se muestra a través de los dashboards hace que las acciones de prevención hayan conducido a la disminución de los accidentes. En particular, la relación entre accidentes y actitudes de empleados fue de gran ayuda para generar acciones preventivas. Se indica también si el accidente ocurrió por falta de capacitación, falta de conocimiento o falta de responsabilidad [16].

A través de la propuesta de planificación general del dominio temático de la inteligencia de negocios de las empresas de energía eléctrica, y la utilización integral del almacén de datos, el análisis y procesamiento en línea y el modelo de análisis matemático, etc., se proporciona una vista personalizada para el personal de gestión y operación personal en varios niveles. De esta forma, se puede esperar que la capacidad de adquirir información de análisis de gestión de las empresas mejore, lo que fomentará aún más la calidad y la velocidad de la gestión y la decisión [17].

En el área de minería de datos, Seon Yeong Han, JaeGoo No, Jin-Ho Shin y YongJae Joo proponen un mecanismo para la detección de anomalías usando datos generados por sistemas AMI que incorpora la probabilidad condicional en la determinación de la normalidad. La novedad de su estudio es generar un espacio bidimensional utilizando la similitud y la probabilidad condicional, por lo que se pueden aplicar varios métodos de clasificación multidimensionales. Ellos comparan el mecanismo propuesto con los métodos de detección de anormalidades basadas en prototipos de mejor ajuste y promedio. En conclusión, el mecanismo propuesto puede distinguir los datos de fraude con una mayor precisión que los métodos tradicionales. También exploran la precisión del mecanismo con varios parámetros[18].

Chen Rui, Hou Yibin, Huang Zhangqin, y He Jian propusieron un modelo de datos, así como un modelo de gestión de datos .ªmI-Data"para el espacio AmI[19]. AmI-Data es compatible con el almacenamiento y la recolección automática de datos de múltiples fuentes de datos heterogéneas. También es compatible con metadatos para dar al sistema la capacidad de descubrimiento de servicios entre sistemas AmI heterogéneos. Además, AmI-Data también es compatible con muchos métodos avanzados de minería de datos, como HMM, ANN, Fuzzy computing. La presentación de datos final de AmI-Data es un "servicio", que es el objetivo final de las aplicaciones de inteligencia ambiental.

Finalmente, Wanrong Qiu, Feng Zhai, Zhejing Bao, Baofeng Li, Qiang Yang, y

Yongfeng Cao investigaron el enfoque de agrupamiento y los índices característicos para los perfiles de carga de los clientes[20]. En primer lugar, se propone un método para extraer el perfil de uso de electricidad típico de un cliente individual, en el cual el algoritmo DBSCAN basado en la densidad es seguido por una estrategia de corrección que apunta a eliminar la distorsión significativa en los resultados de extracción causados por DBSCAN. En segundo lugar, la agrupación de K-means se realiza para agrupar el comportamiento de los perfiles de carga típicos de los clientes y la función de evaluación para el efecto de cluster se aplica para determinar el número más apropiado de clusters. Finalmente, se presentan los índices de caracterización con los que el cluster de perfil de un cliente podría ser identificado fácilmente. La efectividad de los métodos propuestos queda demostrada por los datos reales del consumo de energía de los clientes industriales y comerciales de AMI.

En la actualidad, el enfoque de la construcción de TI se ha desplazado desde la informatización básica de la organización a la optimización de la misma basada en el servicio de información. Por lo tanto, la transformación efectiva de los datos de los sistemas de inteligencia de negocios en el conocimiento que respalda las estrategias empresariales / decisiones comerciales se ha convertido en la orientación de desarrollo de la construcción de informatización de empresas. En Colombia, es clara la apuesta del sector eléctrico colombiano por el desarrollo e implantación de tecnologías que permiten dotar de inteligencia a la red de distribución de servicio de energía eléctrica. Pero aún no se cuentan con sistemas de aprovechamiento y obtención de información por medio de Inteligencia de Negocios y minería de datos.

8. Marco Teórico

Bases de Datos

Desde el auge computacional experimentado en el mundo, las bases de datos han cobrado vital importancia debido a su gran utilidad y amplia implementación en diversas áreas de la sociedad; estas se han convertido en elementos indispensables en las organizaciones debido a la facilidad con la que permiten gestionar información, además, de aportar directamente a la productividad de la organización gracias la rapidez y agilidad que aportan a los diferentes procesos.

"Una base de datos es una colección compartida de datos lógicamente relacionados, y una descripción de esos datos, diseñada para satisfacer las necesidades de información y respaldar las actividades de una organización. Se implementa una base de datos en un sistema de administración de bases de datos (DBMS), que es un sistema de software que permite a los usuarios definir, crear, manipular y administrar una base de datos" [21].

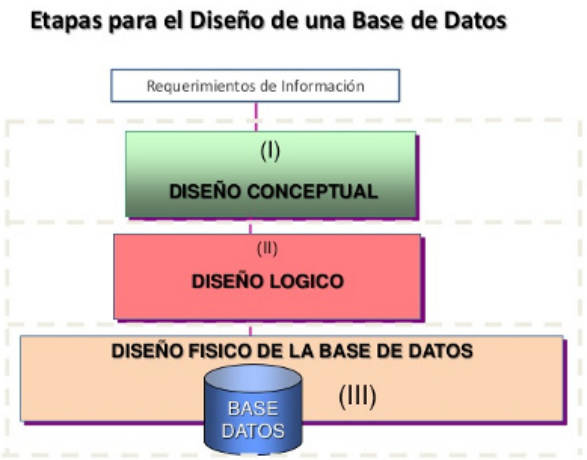


Figura 1: Fases de diseño de una base de datos[1].

Básicamente, con el objetivo de generar independencia en los datos la implementación de una base de datos está dividida en 4 fases (Figura 1) , especificación de requerimientos, diseño conceptual, diseño lógico y diseño físico. La especificación de requerimientos consiste en la recopilación de información de las necesidades del usuario con respecto a la base de datos y la categorización y priorización de las mismas. En el diseño conceptual se realiza un modelo que plasma las entidades y relaciones relevantes, este puede realizarse usando dos enfoques diferentes, diseño de arriba hacia abajo o de abajo hacia arriba. El diseño lógico consiste en la traducción del modelo conceptual a un

modelo lógico más relacionado con el DBMS, este modelo puede ser modelo relacional , orientado a objetos, semi-estructurado, entre otros. Finalmente, en la fase de diseño físico consiste en la implementación del modelo lógico en una plataforma DBMS, los DBMS comunes incluyen SQL Server, Oracle, DB2 y MySQL, entre otros.

Bases de Datos Relacionales

Una relación se define como: "Dado conjuntos S1, S2, ..., Sn (no necesariamente distintos), R es una relación en estos n conjuntos si es un conjunto de n-tuplas, cada uno de los cuales tiene su primer elemento de S1, su segundo elemento de S2, y así sucesivamente" [22]. Así, una base de datos relacional está implementada sobre un modelo relacional que hace uso de la lógica de predicados y teoría de conjuntos y "es una base de datos en la que: los datos son percibidos por el usuario como tablas y los operadores disponibles para su recuperación por el usuario son operadores que derivan nuevas tablas de las antiguas"[23].
Las bases de datos relacionales se caracterizan por tener entidades, atributos y relaciones; son íntegras, es decir los datos tienen mayor validez y consistencia; pueden ser accedidas concurrentemente; ofrecen distintas posibilidades de vistas; cuentan con independencia física y lógica; buen control de información redundante y son normalizadas. Podemos definir la Teoría de la Normalización como la descomposición sin pérdida de información ni de semántica de la relación universal (o de una colección de relaciones equivalentes a la misma) en una colección de relaciones en la que las anomalías de actualización (inserción, borrado y modificación) no existan o sean mínimas[24]. El proceso de normalización se hace necesario para evitar la redundancia de los datos y las inconsistencias, evitar la incapacidad de almacenar ciertos datos, evitar la ambigüedades y pérdida de información, evitar problemas de actualización (anomalías de inserción,borrado y modificación)de los datos en las tablas y para proteger la integridad de los datos.

Minería de Datos

La minería de datos, popularmente conocida como Descubrimiento del conocimiento en bases de datos (KDD), se refiere a la extracción no trivial de información implícita, previamente desconocida y potencialmente útil de los datos en las bases de datos. Si bien la minería de datos y el descubrimiento de conocimiento en bases de datos (o KDD) se tratan con frecuencia como sinónimos, la minería de datos es en realidad parte del proceso de descubrimiento de conocimiento[25]. Las fuentes de datos pueden incluir bases de datos, bodegas de datos, la Web, otros repositorios de información o datos que se transmiten al sistema de forma dinámica.

En principio, la extracción de datos no es específica para un tipo de medio o datos. La extracción de datos debe ser aplicable a cualquier tipo de repositorio de información.

Sin embargo, los algoritmos y los enfoques pueden diferir cuando se aplican a diferentes tipos de datos. De hecho, los desafíos presentados por diferentes tipos de datos varían significativamente. La minería de datos se está utilizando y estudiando para bases de datos, incluidas bases de datos relacionales, bases de datos relacionales de objetos y bases de datos orientadas a objetos, almacenes de datos, bases de datos transaccionales, repositorios no estructurados y semi-estructurados como la WEB, bases de datos avanzadas como bases de datos espaciales, bases de datos multimedia, bases de datos de series temporales y bases de datos textuales, e incluso archivos planos[26].
El proceso de KDD puede describirse en las siguientes fases(Figura 2). Primero, selección de datos; Segundo, preprocesamiento de datos; Tercero, Transformación de datos; Cuarto, extracción de datos; y finalmente, Interpretación de datos.

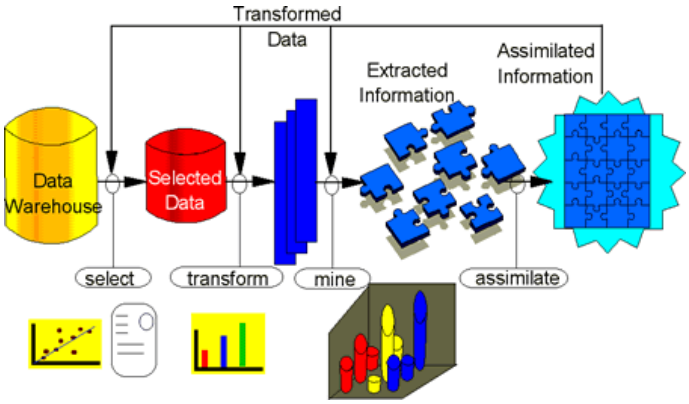


Figura 2: Proceso de minería de datos[2].

BigData

Actualmente, el uso de tecnologías de información ha incrementado de maneras exorbitantes en casi todas las áreas, esto ha generado grandes cantidades de información en formatos difíciles de manipular y a velocidades inimaginables. "Big data es la descripción de una gran cantidad de datos organizados estructuralmente o no que se analizan para tomar una decisión o evaluación informada. Los datos se pueden tomar de una gran variedad de fuentes, incluido el historial de navegación, la geo-localización, las redes sociales, el historial de compras y los registros médicos. Big Data hace referencia a datos complejos que sobrepasan la capacidad de procesamiento de los sistemas de bases de datos simples y tradicionales. Hay tres características principales asociadas con los macro datos: volumen, variedad y velocidad"[27]. Big Data es aplicable a diversas áreas, las áreas de aplicación más conocidas son marketing, computación ubicua e internet de las cosas (IoT).
Para la implementación de un sistema de Big Data se requiere un conjunto de herramientas rico en funcionalidad, como una plataforma de almacenamiento distribuida que pueda mover volúmenes de datos muy grandes al sistema sin perder datos, sistema de configuración para mantener todos los servidores del sistema coordinados, así como for-

mas de encontrar datos y transmitirlos al sistema en algún tipo de flujo basado en ETL y el software también necesita monitorear el sistema y proporcionar sistemas de destino descendentes con feeds de datos para que la administración pueda ver las tendencias y emitir informes en función de los datos.

Hadoop

El proyecto Apache ™ Hadoop® desarrolla software de código abierto para una computación distribuida, confiable y escalable. La biblioteca de software Apache Hadoop es un marco que permite el procesamiento distribuido de grandes conjuntos de datos en clusters de computadoras que usan modelos de programación simples. Está diseñado para escalar desde servidores únicos a miles de máquinas, cada una de las cuales ofrecen cómputo y almacenamiento local. En lugar de confiar en el hardware para ofrecer alta disponibilidad, la biblioteca está diseñada para detectar y manejar fallas en la capa de aplicaciones, por lo que entrega un servicio altamente disponible sobre un grupo de computadoras, cada una de las cuales puede ser propensa a fallas[28].
Las características principales de Hadoop son: elimina dificultades de programación paralela, le permite al usuario distribuir ficheros en nodos, que no son otra cosa que ordenadores con hardware básico, capacidad de ejecución de procesos en paralelo en todo momento, tiene módulos de control para la monitorización de los datos, permite la ejecución de consultas, permite el uso de complementos que facilitan el trabajo, manipulación y seguimiento de toda la información que en él se almacena. (Figura 3)

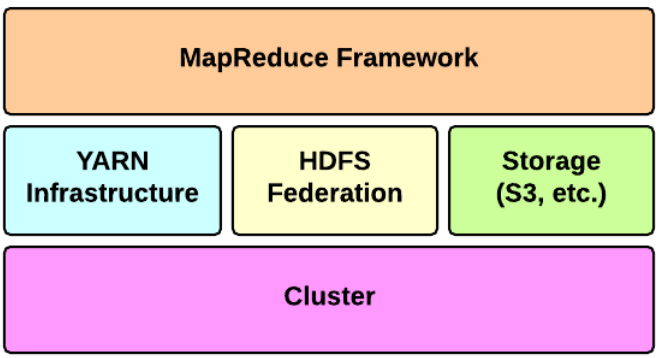


Figura 3: Arquitectura de Hadoop [3].

Inteligencia de Negocios (BI)

En la actualidad gran mayoría de las organizaciones cuentan con información histórica almacenada en las diferentes bases de datos de cada departamento operativo que con el paso de los años deja de ser utilizada en transacciones y procesos diarios, a partir de esta puede obtenerse gran beneficio por medio de la puesta en marcha de sistemas de inteligencia de negocios, que permitan analizar dichos datos y proveer soporte a la

toma de decisiones empresariales.(Figura 4)
"Se entiende por Inteligencia de Negocios al conjunto de metodologías, aplicaciones, prácticas y capacidades enfocadas a la creación y administración de información que permite tomar mejores decisiones a los usuarios de una organización"[29]. Almacenamiento en bodegas de datos, reportes, análisis OLAP, análisis visual, análisis predictivo, cuadro de mando, cuadro de mando integral, minería de datos, gestión de rendimiento, previsiones, reglas de negocios y la integración de datos son algunas de las tecnologías que forman parte de BI.
La BI se caracteriza por la rapidez en la obtención de respuestas, la calidad de la información entregada, versatilidad de manejo (información simple y compleja) y por la fácil interpretación de la información. La implementación de sistemas BI beneficia a una organización por medio de la simplificación del acceso a la información, automatización de la realización de reportes, generación de información unificada y consistente, generalización de la visión de la organización, flexibilización de análisis detección temprana de tendencias, oportunidades y riesgos, y ofreciendo indicadores de rendimiento y desempeño claves para el seguimiento continuo a la organización.

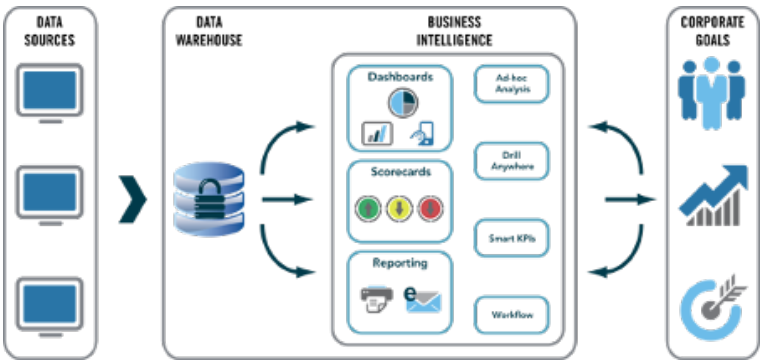


Figura 4: Proceso de BI [4].

Infraestructura de Medición Avanzada (AMI)

Es un paradigma de desarrollo tecnológico, el cual quiere potenciar el desempeño del sistema eléctrico y del suministro de electricidad a partir de las posibilidades que brindan las TIC y establecer nuevos esquemas de participación e interacción de las empresas de suministro de energía con los clientes, con el ciudadano en general. La medición inteligente permite mayores niveles de interactividad de las empresas con sus clientes y de estos con el propio mercado de energía (flujos continuos y bi-direccionales de información) habilitando incluso la participación de los clientes como agentes activos del mercado de energía (ajuste del consumo como respuesta a señales del sistema e incluso flujos bi-direccionales de electricidad). Esto también permite la integración de toda la información del sistema eléctrico operativo (técnica, operativa, financiera, contable, comercial, etc.) y de sub-sistemas diversos y nuevos, heterogéneos, pero inter-

activos lo que haber infinidad de posibilidades tanto para las empresas como para los clientes[30]. AMI integra componentes de software, hardware y arquitecturas y redes de comunicaciones.(Figura 5)

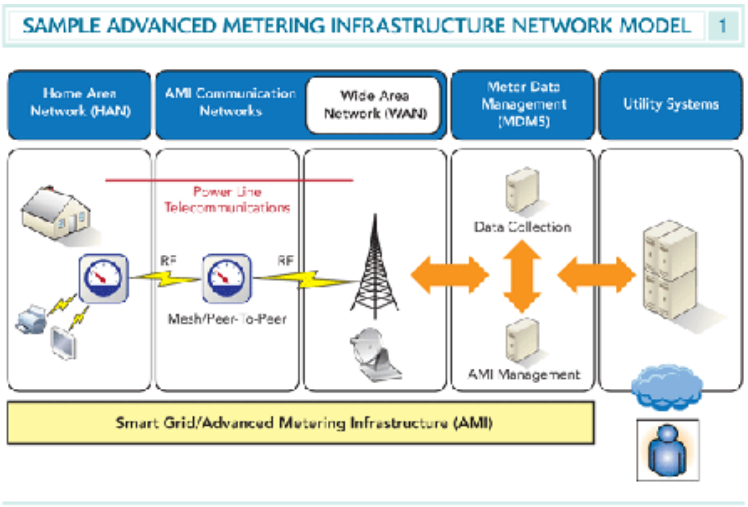


Figura 5: Infraestructura AMI [5].

Parte III
Desarrollo del trabajo

En esta parte del documento se presenta el análisis de herramientas de almacenamiento en bodegas de datos utilizando postgreSQL, el diseño detallado de la bodega de datos, diseño de fases del proceso ETL, descripción y resultados del proceso de implementación de minería de datos, diseño e implementación del módulo de visualización de reportes, plan de pruebas e implementación del mismo con respectivos resultados, plan de despliegue del aplicativo, resultados del trabajo de grado y finalmente conclusiones.

9. Herramientas de almacenamiento de bodegas de datos utilizando PostgreSQL

En esta sección se presenta un estudio detallado de herramientas de gran utilidad para el almacenamiento de grandes volúmenes de datos en bodegas de datos, con el objetivo de establecer criterios de comparación y tener un firme fundamento para elegir cuál es la más apropiada para abordar el tratamiento y por ende un máximo aprovechamiento de los datos. En general, con la implementación de estas herramientas se pretende obtener información con alto grado de veracidad y relevancia que represente el proceso de prestación de servicios eléctricos y pueda ser usada en beneficio de los involucrados en este proceso.

Replicación

La replicación es el proceso que permite la realización de copias de datos entre diversos nodos de almacenamiento, ubicados de forma local o remota, con el objetivo de realizar copias de seguridad o soporte. Las tecnologías de replicación tienen como finalidad garantizar la disponibilidad de los datos y la coherencia de los mismos en diversos nodos de almacenamiento.

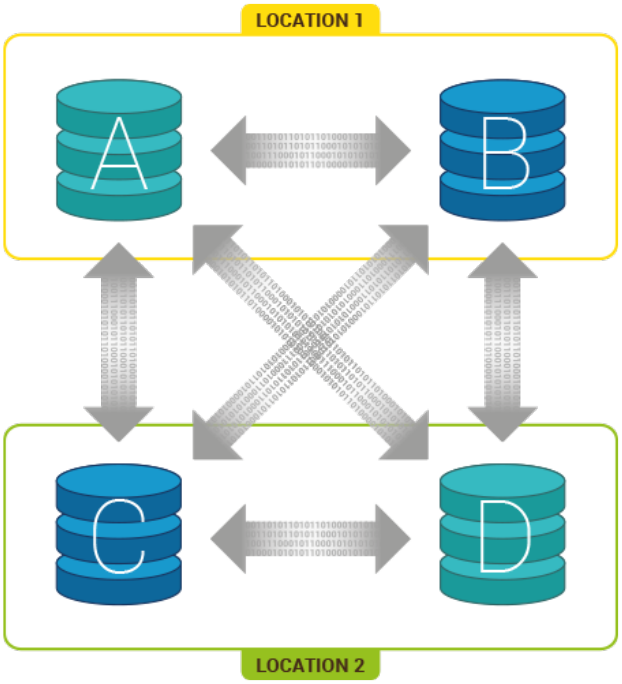


Figura 6: Esquema de nodos distribuidos[6].

Tipos de Replicación

La replicación puede clasificarse con base en tres parámetros: tiempo de latencia, cantidad de maestros que conforman el esquema de replicación y el tipo de dato procesado en la replicación.

1. Según tiempo de latencia
   El tiempo de latencia es el tiempo transcurrido entre la realización de una transacción en un nodo maestro y su replicación en nodos esclavos.
   Replicación síncrona: Las transacciones se replican con tiempos de latencia iguales a cero.
   Replicación asíncrona: Las transacciones se replican con tiempos de latencia mayores a cero.

2. Según cantidad de maestros
   Estructuralmente en la replicación se definen servidores maestros y esclavos. En los maestros se pueden realizar todo tipo de transacciones (Lectura, Escritura y Actualización), Mientras que en los esclavos solo se pueden realizar transacciones de lectura de información.
   Replicación de único maestro : En la replicación de maestro único el esquema de replicación tiene un nodo maestro y todos los nodos restantes son configurados como esclavos, esta configuración es menos flexible que la de maestro múltiple.
   Replicación de múltiple maestro : En la replicación multi-maestro existen en

el esquema de replicación más de un nodo maestro.En caso de que un maestro falle, los demás nodos pueden continuar la escritura y actualización de información, Aunque al existir múltiples maestros se puede presentar pérdidas de consistencia en la información e incrementar la latencia en la comunicación.

3. Según tipo de dato
   Cuando se realiza una transacción en un nodo maestro, se inicia el proceso de replicación de la información, esta información puede enviarse en dos tipos.
   Replicación lógica : La información es replicada como consultas transaccionales que se ejecutan en el nodo esclavo.
   Replicación física : La información es replicada como datos binarios almacenados en los nodos esclavos.

Ventajas

- Aumento en la disponibilidad de la información, debido a que en caso de fallas en un determinado nodo los datos pueden ser obtenidos o gestionados en otro nodo.
- Mejora el rendimiento del sistema gracias a que al aumentar el número de nodos o servidores se pueden distribuir los procesos de respuesta a los usuarios de forma que ninguno de los nodos se sobrecargue aún en casos en los cuales el sistema debe procesar miles de solicitudes en el mismo instante.
- Aumento en la seguridad de la información almacenada debido a que en caso de pérdida de la misma en uno o mas nodos es más probable que otro nodo contenga dicha información perdida y esta se pueda replicar hacia los nodos en los cuales necesita ser recuperada.
- Los costos de implementación son bajos comparados con la implementación de sistemas no distribuidos con el mismo volumen de datos.
- Fácil escalamiento del esquema.

Desventajas

- Aporta complejidad en el software gestor del sistema de almacenamiento distribuido.
- Dependencia de la red de comunicaciones para sincronizar las transacciones sobre cada nodo del sistema, generando sobrecarga en los mismos.
- Costos de mantenimiento incrementan debido a que se deben atender varios subsistemas.
- Dificultad en la manutención de la integridad de los datos debido a que depende de protocolos de comunicación costosos.

- Inexistencia de metodologías y técnicas en el campo de la distribución que actúen como guías en el proceso de implementación y gestión de sistemas distribuidos.
- Complejidad en el diseño del sistema distribuido.

Replicación en PostgreSQL

La Replicación Bidireccional (BDR) es un sistema de replicación asíncrono multi-maestro para PostgreSQL, específicamente diseñado para permitir bases de datos distribuidas geográficamente. BDR,es una extensión de PostgreSQL, libre y de código abierto; licenciada bajo los mismos términos de PostgreSQL.Desarrollado por 2ndQuadrant, BDR está diseñado específicamente para su uso en nodos distribuidos geográficamente, utilizando una replicación lógica asincróna que admite desde 2 hasta más de 48 nodos en bases de datos distribuidas.

Fragmentación

Es una técnica de distribución y escalamiento de bases de datos que consiste en la partición de información alojada en bases de datos en diferentes nodos o servidores con el objetivo de facilitar la gestión de la misma. Al implementar esta técnica se realizan particiones horizontales o verticales sobre tablas de datos con gran volumen de información de forma que se pueda acceder a cada partición por medio de de claves de particionado que permitan ubicar fácilmente el bloque de datos o la partición que contiene la información necesaria a gestionar, disminuyendo así los tiempos de consultas y mejorando el rendimiento del sistema de información. Es importante resaltar que para la correcta implementación de la fragmentación en sistemas de bases de datos se debe cumplir con tres características: Primero, la fragmentación debe ser completa, es decir, cada elemento almacenado en el esquema inicial debe quedar almacenado en algún fragmento del esquema final; segundo, el esquema inicial de datos debe poderse construir a partir del esquema final; y tercero, en el esquema resultante del proceso de fragmentación todos los fragmentos deben ser disjuntos, es decir, la información almacenada en un fragmento de datos no debe estar almacenada en otro fragmento del sistema de almacenamiento.

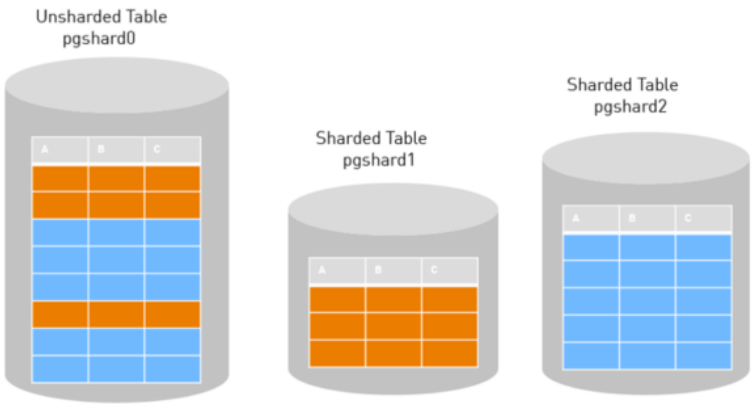


Figura 7: Esquema de partición [7].

Existen dos elementos de gran importancia en el proceso de diseño e implementación de la fragmentación en un esquema de bases de datos que deben ser tenidos en cuenta por el diseñador de dicho esquema debido a que de estos depende en gran manera el éxito de este proceso. Primero, la función de partición que se encarga de la distribución de los datos en los diferentes nodos que conforman el esquema; y segundo, la clave de partición, también conocida como clave de fragmentación o de distribución, que es el conjunto de atributos de los datos sobre los cuales se aplica la función de partición para la correcta distribución de los datos. La elección de la clave y función de partición dependen de la cantidad de nodos que se desea asociar en el esquema, los procesos a optimizar prioritariamente (consulta, inserción, modificación u otros) y la capacidad de gestión de cada nodo a implementar, en caso que estos tengan diferentes características

de rendimiento.

Tipos de Fragmentación

1. Fragmentación Horizontal: La fragmentación se realiza sobre los registros de la tabla.Cada fragmento resultante es un subconjunto de los registros totales de la tabla de origen.Existen dos tipos de fragmentación horizontal: primaria y derivada; En la fragmentación horizontal primaria existen atributos semejantes en cada subconjunto derivado después del proceso de fragmentación. En la fragmentación horizontal derivada se realiza la fragmentación de registros basándose en condiciones definidas sobre atributos de otra relación asociada.

2. Fragmentación Vertical: La fragmentación se realiza con base en los atributos de la tabla y cada fragmento resultante contiene la llave primaria además de algunos atributos de la tabla de origen.

3. Fragmentación mixta: Consiste en la implementación de técnicas de fragmentación vertical y horizontal sobre la misma tabla.

Ventajas

- Disminución en los costos de almacenamiento debido a que se eliminan los duplicados y solo existe una copia de los datos almacenados en el esquema.
- Mejoramiento en la capacidad de concurrencia de acceso a los datos gracias a que se pueden gestionar el mismo tipo de información en diferentes nodos.
- Mejoramiento considerable en el rendimiento de la capa software que hace uso de un fragmento del sistema de almacenamiento.
- Disminución considerable en los tiempos de consultas condicionados directamente con las claves de partición, debido a que se conoce previamente la ubicación de los datos asociados a determinados rangos y/o valores de estas.

Desventajas

- Falta de disponibilidad completa de los datos en casos de fallos de uno o mas nodos.
- La fiabilidad de los datos no se garantiza en los esquemas fragmentados.
- Mayor complejidad en consultas que requieran información situada en diversos nodos del sistema.
- Disminución en el rendimiento del software que hace uso varios fragmentos del sistema de almacenamiento al tiempo.

Fragmentación en PostgreSQL

En PostgreSQL es posible implementar sistemas basados en un esquema de datos fragmentado ya sea por el uso de funcionalidades propias o desarrolladas por terceros, en este caso, usaremos las funcionalidades propias de PostgreSQL. Este esquema puede ser implementado con nodos locales o remotos, en este ultimo caso es necesario usar la extensión de contenedores de datos remotos(FDW) para PostgreSQL que permite la gestión integrada de datos almacenados en diferentes servidores PostgreSQL.

Comparación de Herramientas

Como criterios de evaluación de las herramientas estudiadas se tendrá en cuenta:localización de los datos, fiabilidad de los datos, disponibilidad de los datos, capacidad y costo de almacenamiento, distribución de la carga de procesamiento y costo de comunicación. Se analizaran cada uno de estos criterios y el cumplimiento de los mismos en cada modelo de distribución de datos [31]. Para el proceso de selección de la herramienta a usar, se asignó un valor de relevancia a cada característica, siendo uno (1) de mayor y seis(6) menor importancia para la solución del problema a tratar. Además, se calificó el cumplimiento de cada característica en las herramientas estudiadas como alto, medio o bajo.

Característica	**Relevancia**	**Replicación**	**Fragmentación**
Localización distribuida	1	Si	Si
Fiabilidad	6	Alta	Baja
Disponibilidad	5	Alta	Baja
Capacidad y costo de almacenamiento	2	Alto	Bajo
Distribución de la carga de procesamiento	3	Bajo	Alto
Costo de comunicación	4	Alto	Alto

Tabla 2: Evaluación de Herramientas de Almacenamiento en bodegas de Datos.

Teniendo en cuenta que en el tratamiento y análisis de los datos generados por AMI es de gran relevancia la utilización de tecnologías de distribución de bases de datos debido al gran de volumen de datos procesados, y dando prioridad a la disponibilidad, capacidad de procesamiento y disminución en el costo de comunicación como características del sistema a implementar, se eligió usar un esquema de bases de datos fragmentada. Este esquema además de proveer ventajas como la notable disminución en espacio de almacenamiento requiere menor tiempo y costo de implementación y mayor facilidad en el procesamiento de los datos. En la sección 10 del documento se detalla el esquema final implementado con tecnologías de distribución de bases de datos fragmentadas.

10. Diseño de bodega de datos

Comprensión de los datos

Los datos empleados para el proceso de implementación fueron recolectados en el marco del proyecto Low Carbon London, liderado por UK Power Networks en el Reino Unido entre noviembre de 2011 y febrero de 2014 con el objetivo de reducir las emisiones de dióxido de carbono y mejorar la seguridad en los procesos involucrados en el suministro de energía eléctrica. UK Power Networks recopiló los datos de consumo de energía eléctrica para una muestra representativa de 5,567 hogares de Londres tomando lecturas cada media hora usando sistemas AMI. Dentro del conjunto de datos hay dos grupos de clientes. El primero esta conformado por 1,100 clientes sujetos a tarifas de Tiempo de uso dinámico (ToU) a los cuales se les asignó un horario específico cuando su tarifa de consumo de energía eléctrica sería alta (67.20p / kWh), baja (3.99p / kWh) o normal (11.76p / kWh). Los precios de las tarifas se entregaron un día antes a través de la pantalla de inicio del medidor inteligente instalado en el hogar o por medio de mensajes de texto al teléfono móvil del cliente. El resto de la muestra, alrededor de 4,500 clientes tenía una tarifa de consumo de energía eléctrica fija de 14.228 p / kWh.

Todos los datos están asociados a grupos y categorías ACORN como parte de la clasificación demográfica. ACORN es una poderosa clasificación de consumidores que segmenta a la población del Reino Unido. Al analizar los datos demográficos, los factores sociales, la población y el comportamiento del consumidor, proporciona información precisa y una comprensión de los diferentes tipos de personas. Acorn proporciona información valiosa para el consumidor que lo ayuda a orientarse, adquirir y desarrollar relaciones rentables con los clientes y mejorar la prestación de servicios.[8]

Category 1 Affluent Achievers	**Group A: Lavish Lifestyles**
	Group B: Executive Wealth
	Group C: Mature Money
Category 2 Rising Prosperity	**Group D: City Sophisticates**
	Group E: Career Climbers
Category 3 Comfortable Communities	**Group F: Countryside Communities**
	Group G: Successful Suburbs
	Group H: Steady Neighbourhoods
	Group I: Comfortable Seniors
	Group J: Starting Out
Category 4 Financially Stretched	**Group K: Student Life**
	Group L: Modest Means
	Group M: Striving Families
	Group N: Poorer Pensioners
Category 5 Urban Adversity	**Group O: Young Hardship**
	Group P: Struggling Estates
	Group Q: Difficult Circumstances

Figura 8: Clasificación Acorn [8].

Requerimientos

La bodega de datos deberá estructurarse de forma tal que permita la obtención de información de consumo eléctrico. se deben cumplir los siguientes requisitos:

- Se permitirá la consulta del consumo eléctrico categorizada según fechas por días, semanas, meses, trimestres, estaciones, fines de semana y días festivos
- Se permitirá la consulta del consumo eléctrico categorizada según los tipos de facturación(estándar y dinámica).
- Se permitirá la obtención de información de los usuarios basados en la categorización según los grupos ACORN
- Se permitirá la consulta del consumo eléctrico categorizada según las horas del día y las diferentes franjas horarias
- Se permitirá la consulta del consumo eléctrico para tarifas dinámicas según el tipo de tarifa(bajo, normal y alto)

Diseño Conceptual

Para cumplir con los requerimientos planteados se diseñaron tres(3) datamarts modelados bajo el diseño de estrella; Los datamarts contienen información de consumo

y costo del servicio de energía eléctrica de usuarios residenciales con granularidad de franjas, días y meses respectivamente.

Dimensiones

Se diseñaron las siguientes dimensiones con el objetivo de parametrizar los datos en las tablas de hechos de los tres datamarts definidos.

Nombre	**Descripción**
Día	Días en los que se tomaron medidas de consumo.
Mes	Meses en los que se tomaron medidas de consumo.
Franja	Contiene información de la agrupación de horas del día.
Cliente	Esta dimensión contiene la información del cliente.
Grupo	Esta dimensión contiene información demográfica asignada a los clientes según los grupos definidos en ACORN
Categoría	Esta dimensión contiene información demográfica de los clientes basada en el agrupamiento de grupos ACORN.

Tabla 3: Descripción de dimensiones.

Las dimensiones descritas se relacionan de la siguiente forma con cada datamart:

Dimensiones / **DataMarts**	Franja	Día	Mes	Cliente	Grupo	Categoría
Consumo Por Franjas	X	X		X	X	X
Consumo Diario		X		X	X	X
Consumo Mensual			X	X	X	X

Tabla 4: Matriz de DataMarts

Datamart consumo franjas

Contiene la información de consumos agrupada según la franja horarias para cada cliente.

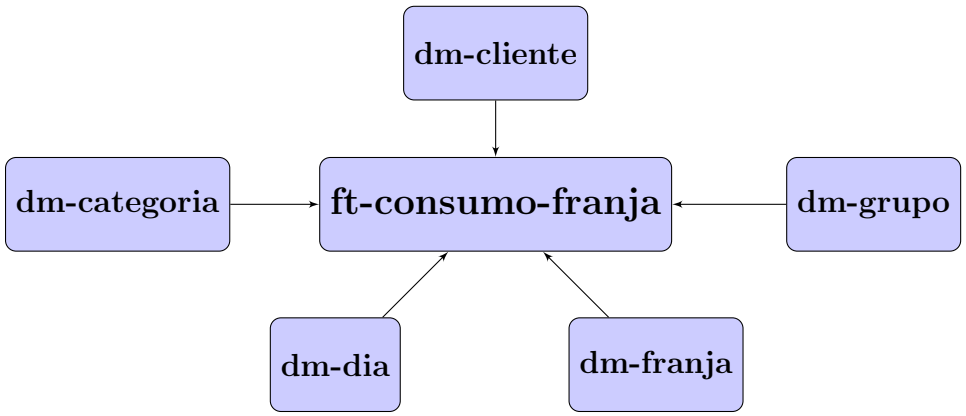


Figura 9: Datamart consumo franjas.

Datamart consumo diario

Contiene la información de consumos agrupada según la día para cada cliente.

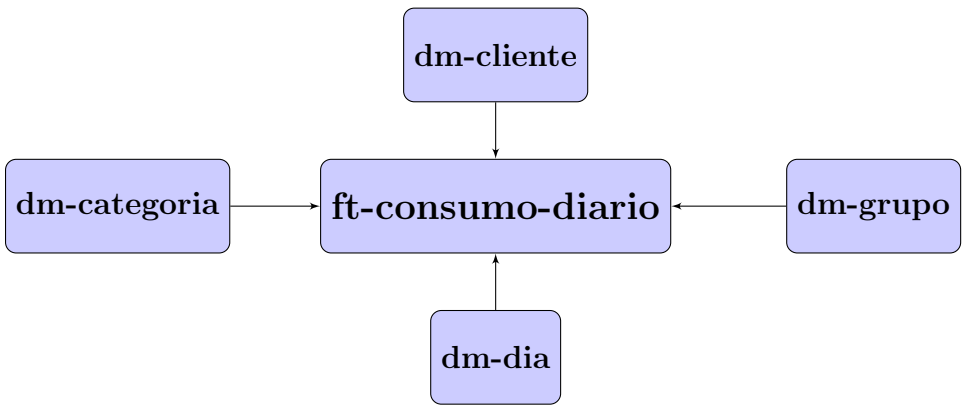


Figura 10: Datamart consumo diario.

Datamart consumo mensual

Contiene la información de consumos agrupada según el mes fecha para cada cliente.

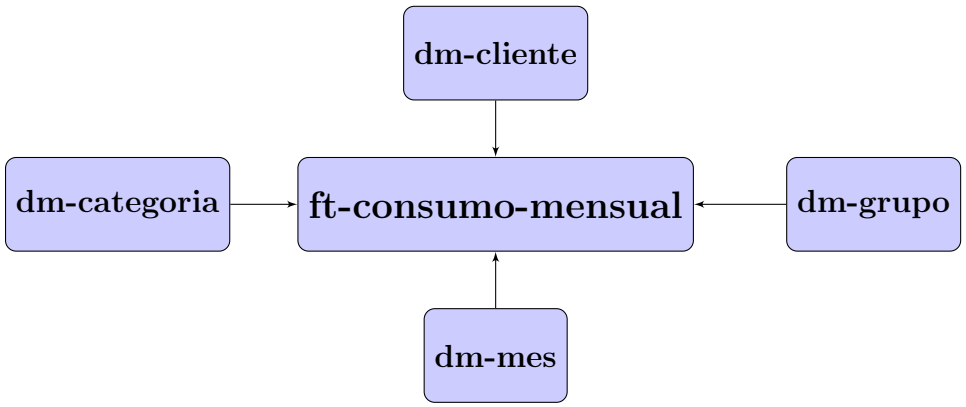


Figura 11: Datamart consumo mensual.

Diseño Conceptual Fragmentado

Se definió la aplicación de fragmentación de tipo horizontal sobre las tablas de hechos de los 3 datamarts de consumo(franjas, diario, mensual) con el objetivo de generar 2 nodos, seleccionando como clave de fragmentación el ID del cliente del servicio de energía eléctrica, en el Nodo 1 se almacenaran los registros de consumos que tengan asociado un cliente con identificador impar, mientras en el Nodo 2 se almacenaran los consumos que estén asociados a clientes con identificador par.

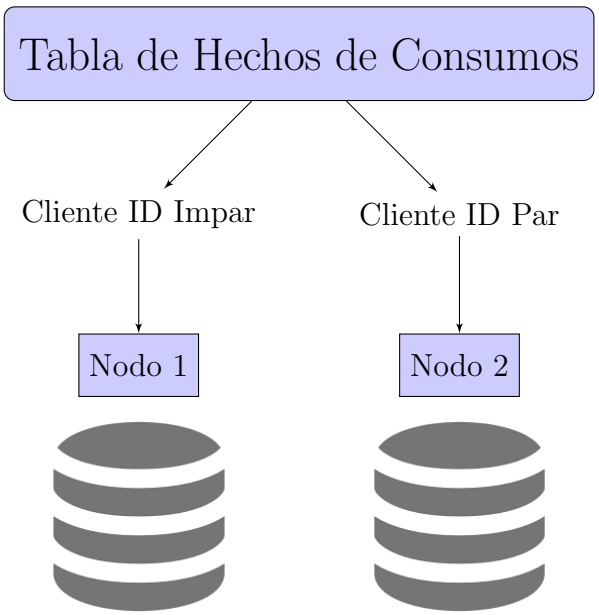


Figura 12: Esquema de fragmentación de tablas de hechos.

11. Módulo ETL

El módulo de proceso ETL es esencial para cumplir con los objetivos planteados debido a que este genera la información base para los procesos de inteligencia de negocios y minería de datos. Este proceso está conformado por tres fases principales: extracción, transformación y carga. Para la implementación del proceso ETL se utilizó Pentaho Data Integration(PDI) que es una herramienta perteneciente a la suite libre Pentaho orientadas a solución de problemas de Inteligencia de Negocios(BI), PDI permite implementar los tres procesos de ETL con diversas opciones de configuraciones e integración con varias herramientas de lectura, procesamiento y almacenamiento de información. En este caso, se realizó el ETL en cuatro fases: ETL de dimensiones, precarga de datos de hechos, agrupamiento de consumos por franjas, agrupamiento de consumos por días y agrupamiento de consumos por meses; en todas estas fases se involucraron las tres etapas del proceso ETL.

Requerimientos

El objetivo general del proceso ETL es tomar las fuentes de datos entregadas para la implementación del proyecto y realizar operaciones de extracción, transformación, y carga de datos en la bodega de datos diseñada.Específicamente se definieron los siguientes requerimientos:

- Realizar la carga de datos a partir de archivos delimitados.
- Realizar la generación automática de los datos de las dimensiones temporales días y meses.
- Cargar datos de consumo en tablas de hechos agrupadas por franjas, días y meses.
- Calcular consumos y costos totales para cada tabla de hechos según la fecha y hora del consumo en los 3 datamarts definidos.
- Estandarizar formatos de fechas y horas en las diferentes dimensiones.
- Relacionar información contenida en dimensiones y tablas de hechos mediante llaves primarias de cada registro.

Extracción

En el proceso de extracción se realizó la lectura de las fuentes de datos. Se tomaron fuentes de información en formato de archivos de texto con extensión .csv delimitados por comas. En la tabla 5 se describen los archivos fuentes y la información extraída de cada uno.Adicionalmente, se construyeron procesos para el llenado de las dimensiones días y meses. Después de la lectura de datos de las fuentes se procedió a almacenar en tablas temporales todos los datos que necesitaran la aplicación de transformaciones.

Archivo Fuente	Descripción
categorias.csv	Contiene el id y el nombre de las 5 categorías definidas en la clasificación demográfica ACORN[8].
consumos.csv	Contiene los datos de consumo para cada hogar participante en el proyecto Low Carbon London entre noviembre de 2011 y febrero de 2014; se detalla identificador del hogar, el tipo de energía(ToU o STD), la fecha y la hora ,consumo en kWh(por cada media hora), grupo ACORN del hogar ,categoría ACORN del hogar.
festivos.csv	Contiene los datos de los días feriados en Londres en el periodo de los datos de consumo. Se detalla la fecha y el tipo de día feriado.
franjas.csv	Contiene los datos de las 8 franjas horarias diarias elegidas para el análisis de los datos. Se detalla el id de la franja, la hora de inicio y hora de finalización.
grupos.csv	Contiene la información de los grupos definidos en la clasificación demográfica de ACORN. Se detalla el identificador, acrónimo, nombre y categoría a la que pertenece el grupo.
tarifas.csv	Contiene las tarifas del servicio de energía aplicables a clientes sujetos a tarifas ToU según la fecha y la hora. Estas tarifas se clasifican en bajas, normales o medias. Se detalla la fecha, hora y tipo de tarifa.

Tabla 5: Fuentes de datos

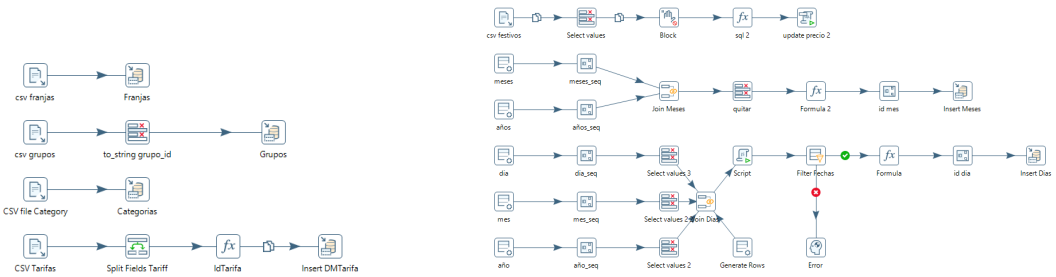
Transformación

Después de extraer la información se aplicó el proceso de transformación de los datos que incluyó cambios de tipos de datos en algunos atributos que debido a que las fuente eran textos planos estaba representados inicialmente como datos alfanuméricos y debían ser procesados como datos numéricos; se separó información que estaba representada como única, pero por la naturaleza del proceso proveía varios datos de utilidad en el análisis (fechas y horas); se realizó el cálculo de los consumos y costos según las horas, fechas y tipos de tarifas establecidas para cada cliente; se estandarizó el formato de las fechas provenientes de todas las fuentes de datos y las generadas automáticamente; se indexó la información de las tablas de hechos tomando como referencia la información de las dimensiones; Finalmente, se eliminó la información no valiosa para el proceso.

Carga

En el proceso de carga se tomó la información procesada y se almacenó en la bodega de datos montada en PostgreSQL con la estructura definida en el diseño dimensional.

Diagramas del Proceso ETL



(a) Carga de fuentes de datos. (b) Generación automática.

Figura 13: ETL dimensiones.

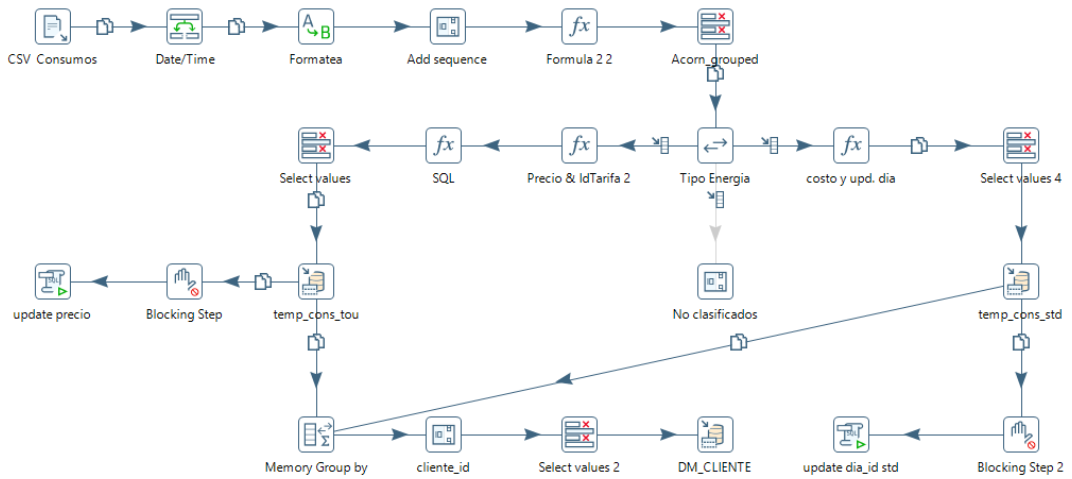
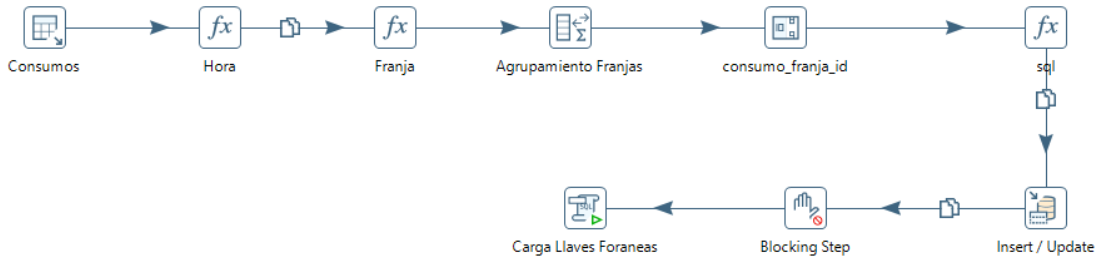
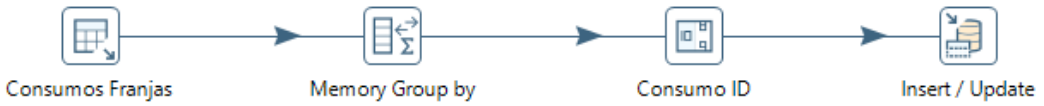


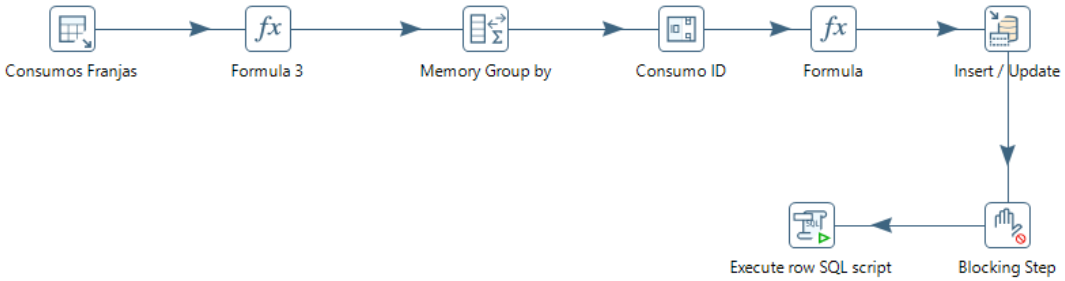
Figura 14: ETL de hechos(Precarga).



(a) Agrupación hechos por franjas.



(b) Agrupación hechos por días.



(c) Agrupación hechos por meses.

Figura 15: ETL de hechos(Agrupacion).

12. Minería de Datos

En esta sección se detalla la implementación de minería de datos al conjunto de datos previamente definido para obtener conocimiento sobre los patrones de consumo de energía eléctrica similares entre los diferentes grupos y categorías ACORN de los usuarios. Según lo definido previamente, la minería de datos es un grupo de tecnologías y técnicas aplicables a grandes cantidades de datos con el objetivo de revelar conocimiento de gran valor para la organización a partir de estos. La selección de la técnica de minería de datos a usar depende del objetivo del negocio, debido a que no existe una mejor técnica que otra sino que existen técnicas que aplicadas a determinados problemas ofrecen mejores resultados que otras.

Clustering

El clustering es una de las técnicas mas importantes de minería de datos y aprendizaje no supervisado, consiste en la aplicación de algoritmos sobre grandes cantidades de datos para agruparlos en clusters o grupos según las características de los datos y cumpliendo las siguientes reglas: primero, los elementos en un mismo grupo deben ser lo mas similares posibles; segundo, los elementos en diferentes grupos deben ser lo mas diferentes posibles y tercero, la medición de la similitud y disimilitud debe ser clara y tener un significado practico [32].

Selección de Datos

El proceso de minería de datos se implementara sobre las tablas de hechos de los datamarts de consumos por franjas, diarios y mensuales con el propósito de comparar la correlación existente entre los consumos y las características demográficas(grupos y categorías Acorn) previamente definidas en cada usuario. Las variables influyentes en las tablas de hechos de las bodegas de datos de consumos son las siguientes:

Consumo Franjas	**Consumo Diario**	**Consumo Mensual**
Cliente	Cliente	Cliente
Consumo	Consumo	Consumo
Costo	Costo	Costo
Franja Horaria	Día de la Semana	Mes del Año

Tabla 6: Variables Influyentes en el Modelo de Minería de Datos.

Implementación

Se eligió la aplicación del algoritmo K-Means[33] por medio de la aplicación WEKA (Waikato Environment for Knowledge Analysis)[34] un software de código abierto y con Licencia Pública General(GNU) que ofrece una colección de algoritmos de aprendizaje

automático y minería de datos, estos algoritmos permiten la ejecución de tareas de preprocesamiento de datos y además la implementación de técnicas como agrupación, clasificación, regresión, entre otras. La versión de WEKA usada es la 3.8.3. Una de las desventajas del uso del algoritmo K-Means para el agrupamiento de instancias, es que se debe definir el valor de K(número de grupos resultantes) antes de ejecución del algoritmo, un método apropiado para la definición del valor óptimo del parámetro K es el método del codo[35] que consiste en la ejecución del algoritmo K-Means con un determinado rango de valores de K, para cada valor se calcula la suma de errores al cuadrado (SSE) y se traza una gráfica de lineas de los valores de SSE en función de los valores de K. Finalmente, se analiza la gráfica con el objetivo de encontrar del punto de quiebre(codo) y se define como el valor óptimo de grupos el K correspondiente al punto de quiebre.

El procedimiento experimental consiste en la ejecución del algoritmo K-Means con K=5 y K=17 con el objetivo de realizar la comparación del agrupamiento resultante con las 5 categorías y 17 grupos Acorn asociados previamente a los usuarios del servicio de energía eléctrica residencial. Además de encontrar el valor K óptimo, ejecutar el algoritmo con este valor y realizar una comparación de los grupos resultantes de ejecutar el algoritmo K-Means sobre los consumos por franjas, días y meses con los tres valores de K.

Agrupación de Consumos por Franjas

Se aplico el algoritmo K-Means a un conjunto de datos de 44.494 consumos promedios de energía eléctrica por franjas horaria de los usuarios del servicio. Cada instancia tiene como atributos de análisis el identificador del cliente, franja horaria, el consumo promedio y costo promedio; adicionalmente dos atributos que permiten analizar la correcta agrupación de los consumos: el grupo y la categoría Acorn.

Agrupación de Consumos por Días

Se aplico el algoritmo K-Means a un conjunto de datos de 68.330 consumos promedios de energía eléctrica por días de la semana de los usuarios del servicio. Cada instancia tiene como atributos de análisis el identificador del cliente, día de la semana, el consumo promedio y costo promedio; adicionalmente dos atributos que permiten analizar la correcta agrupación de los consumos: el grupo y la categoría Acorn.

Agrupación de Consumos por Meses

Se aplico el algoritmo K-Means a un conjunto de datos de 65.686 consumos promedios de energía eléctrica por meses del año de los usuarios del servicio. Cada instancia tiene como atributos de análisis el identificador del cliente, mes del año , el consumo promedio y costo promedio; adicionalmente dos atributos que permiten analizar la correcta agrupación de los consumos: el grupo y la categoría Acorn.

Resultados Experimentales

Selección del Numero Óptimo de Clusters

Se aplico el método del codo para un numero de clusters K entre 2 y 17, en la Figura 16 se visualizan los resultados obtenidos. Para los consumos por franjas, días y meses se puede observar un cambio brusco en los valores de SSE para K = 4, lo cual nos indica que este el numero óptimo de clusters a parametrizar en el algoritmo K-Means.

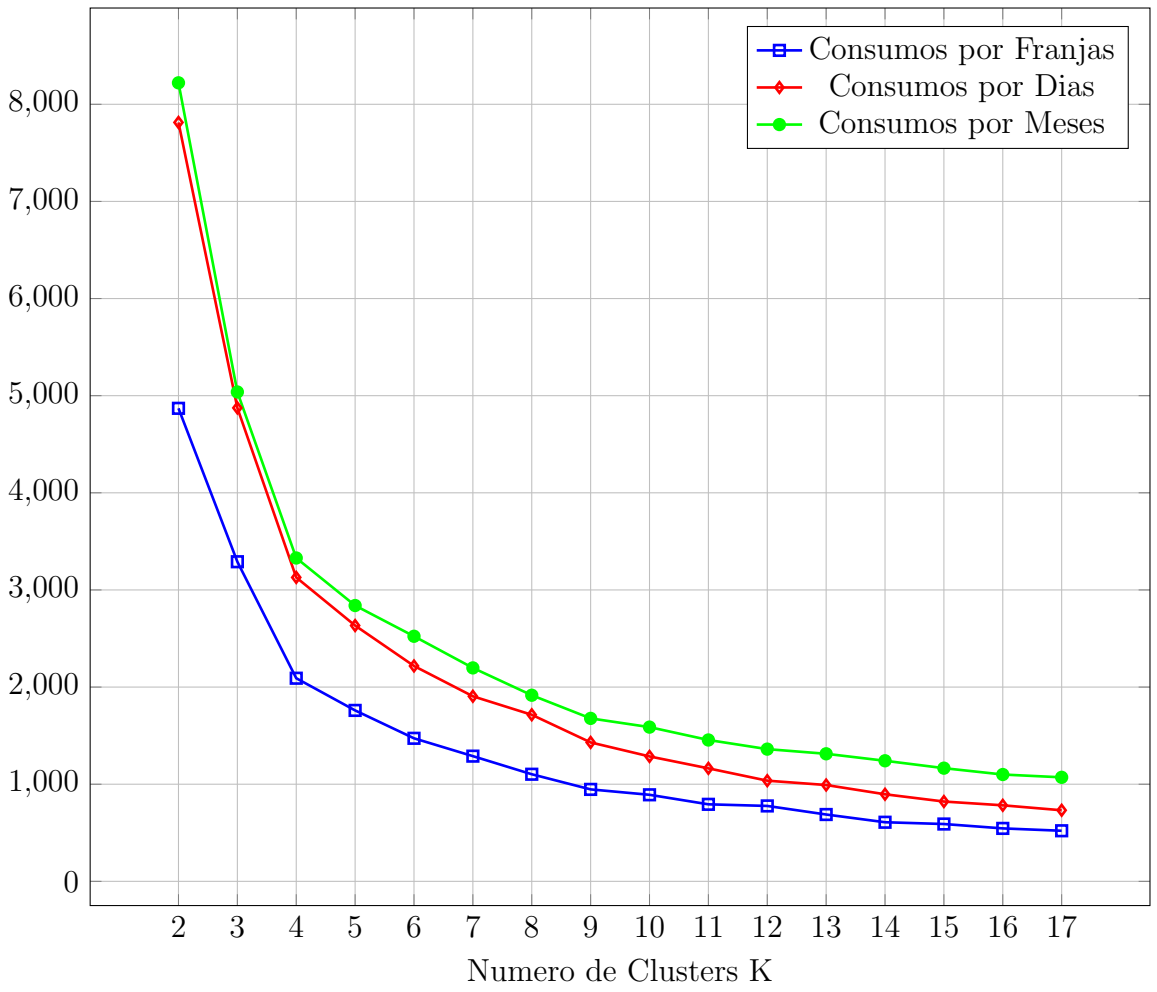


Figura 16: Método del codo.

Agrupación con K = 4 (K Óptimo)

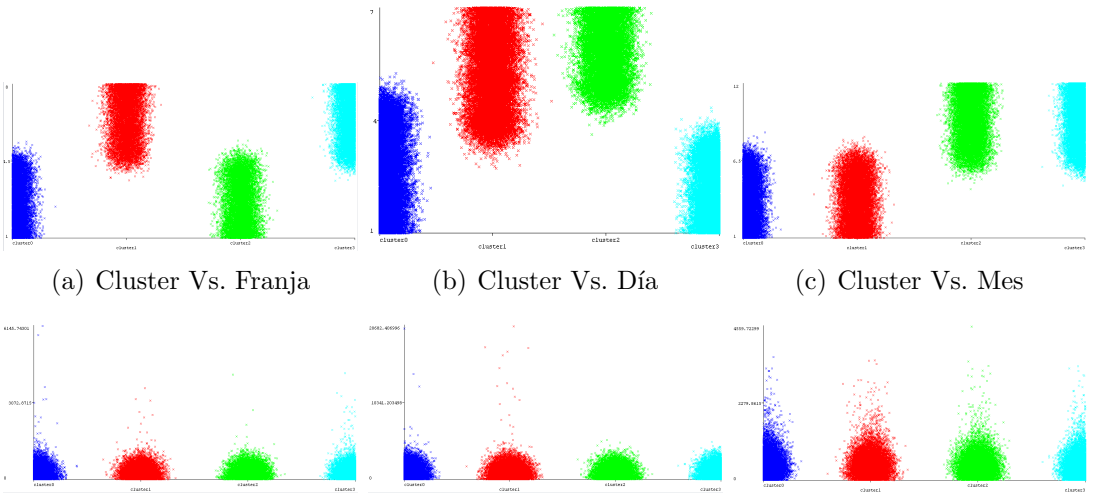
En esta sección se analizan los resultados de aplicar el algoritmo K-Means predefiniendo cuatro clusters como valor de K sobre los tres conjuntos de datos de consumos que son consumos agrupados por franjas, días y meses. En la Tabla 7 se detallan las cantidades de instancias de datos asociadas a cada uno de los 4 clusters y su rela-

ción porcentual con respecto al total de datos de la muestra, se puede observar una agrupación de datos porcentualmente uniforme en todos los casos.

#	Franjas		Días		Meses	
	Cantidad	**Porcentaje**	**Cantidad**	**Porcentaje**	**Cantidad**	**Porcentaje**
1	11128	25 %	20286	30 %	16393	25 %
2	11125	25 %	18911	28 %	16317	25 %
3	11116	25 %	14160	21 %	16453	25 %
4	11124	25 %	14973	22 %	16523	25 %

Tabla 7: Instancias agrupadas en cada grupo con K=4.

En las gráficas a, b, y c de la Figura 17 se puede visualizar la relación existente entre los respectivos atributos temporales(franja horaria, día de la semana y mes del año) y los clusters resultantes. Y en las gráficas d, e y f de la Figura 17 se puede visualizar la relación existente entre el consumo promedio para dichos atributos temporales y los clusters resultantes.



(a) Cluster Vs. Franja (b) Cluster Vs. Día (c) Cluster Vs. Mes

(d) Cluster Vs. Cons. Franja (e) Cluster Vs. Cons. Día (f) Cluster Vs. Cons. Mes

Figura 17: Clusters de consumos para K=4.

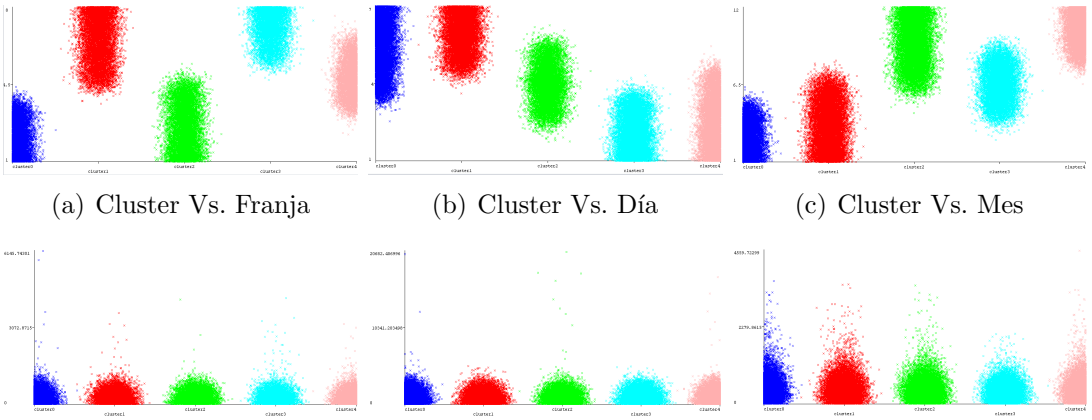
Agrupación con K = 5

En esta sección se analizan los resultados de aplicar el algoritmo K-Means predefiniendo cinco clusters como valor de K sobre los tres conjuntos de datos de consumos que son consumos agrupados por franjas, días y meses. En la Tabla 8 se detallan las cantidades de instancias de datos asociadas a cada uno de los 4 clusters y su relación porcentual con respecto al total de datos de la muestra.

#	Franjas		Días		Meses	
	Cantidad	**Porcentaje**	**Cantidad**	**Porcentaje**	**Cantidad**	**Porcentaje**
0	8610	19 %	13334	20 %	11319	17 %
1	10444	23 %	10166	15 %	15424	23 %
2	10426	23 %	12181	18 %	15525	24 %
3	7442	17 %	16574	24 %	11956	18 %
4	7571	17 %	16075	24 %	11462	17 %

Tabla 8: Instancias agrupadas en cada grupo con K=5.

En las gráficas a, b, y c de la Figura 18 se puede visualizar la relación existente entre los respectivos atributos temporales(franja horaria, día de la semana y mes del año) y los clusters resultantes. Y en las gráficas d, e y f de la Figura 18 se puede visualizar la relación existente entre el consumo promedio para dichos atributos temporales y los clusters resultantes.



(a) Cluster Vs. Franja (b) Cluster Vs. Día (c) Cluster Vs. Mes

(d) Cluster Vs. Cons. Franja (e) Cluster Vs. Cons. Dia (f) Cluster Vs. Cons. Mes

Figura 18: Clusters de consumos para K=5.

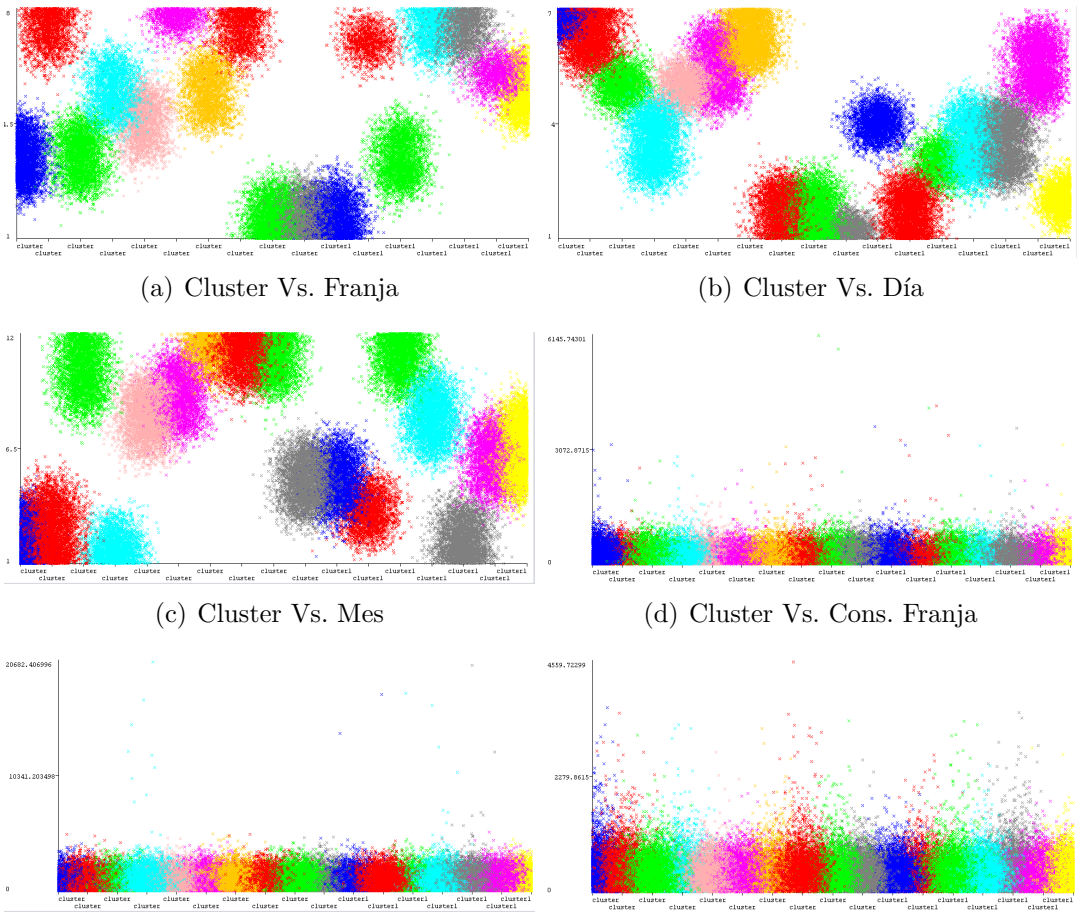
Agrupación con K = 17

En esta sección se analizan los resultados de aplicar el algoritmo K-Means predefiniendo cinco clusters como valor de K sobre los diecisiete conjuntos de datos de consumos que son consumos agrupados por franjas, días y meses. En la Tabla 9 se detallan las cantidades de instancias de datos asociadas a cada uno de los 4 clusters y su relación porcentual con respecto al total de datos de la muestras.

#	Franjas		Días		Meses	
	Cantidad	**Porcentaje**	**Cantidad**	**Porcentaje**	**Cantidad**	**Porcentaje**
0	3384	8 %	2215	3 %	4090	6 %
1	2120	5 %	4679	7 %	4733	7 %
2	3386	8 %	3033	4 %	3768	6 %
3	2641	6 %	5231	8 %	2984	5 %
4	2536	6 %	3036	4 %	4335	7 %
5	1322	3 %	4929	7 %	4046	6 %
6	2757	6 %	4688	7 %	2545	4 %
7	2116	5 %	3682	5 %	3195	5 %
8	3706	8 %	4961	7 %	3465	5 %
9	3712	8 %	2644	4 %	4235	6 %
10	3704	8 %	3193	5 %	4274	7 %
11	1322	3 %	4566	7 %	3112	5 %
12	3259	7 %	3199	5 %	3470	5 %
13	2120	5 %	5242	8 %	4365	7 %
14	2122	5 %	5180	8 %	4078	6 %
15	1534	3 %	4855	7 %	4359	7 %
16	2752	6 %	2997	4 %	4632	7 %

Tabla 9: Instancias agrupadas en cada grupo con K=17.

En las gráficas a, b, y c de la Figura 19 se puede visualizar la relación existente entre los respectivos atributos temporales(franja horaria, día de la semana y mes del año) y los clusters resultantes. Y en las gráficas d, e y f de la Figura 19 se puede visualizar la relación existente entre el consumo promedio para dichos atributos temporales y los clusters resultantes.



(a) Cluster Vs. Franja

(b) Cluster Vs. Día

(c) Cluster Vs. Mes

(d) Cluster Vs. Cons. Franja

(e) Cluster Vs. Cons. Dia

(f) Cluster Vs. Cons. Mes

Figura 19: Clusters de consumos para K=17.

Conclusiones

A partir de los diagramas de clusters presentados en las Figuras 17, 18, 19 se puede concluir:

- Los atributos temporales(franja, día, mes) tienen una influencia alta en la agrupación resultante de aplicar el algoritmo K-Means con 4, 5 y 17 clusters respectivamente.
- El atributo consumo no representa una gran influencia en el modelo de agrupación resultante de la aplicación de K-Means con 4, 5 y 17 clusters respectivamente.
- Los resultados obtenidos no revelan la existencia de patrones de consumo de energía eléctrica de los usuarios, por tanto, debido a la irrelevancia de los datos resultantes, no se incluye esta información en el modulo de visualización de reportes.

13. Módulo de Visualización de Reportes

En esta sección se detalla el proceso de diseño e implementación del módulo de visualización de reportes.

Requerimientos

Requerimientos funcionales

El objetivo del módulo es permitir la interacción del usuario con la información generada en los procesos de inteligencia de negocios y minería de datos por medio de reportes gráficos que faciliten la comprensión de la misma; además la información debe filtrarse por:

RF1. Visualización de gráficos de datos según grupos Acorn, categorías Acorn o tipos de tarifa.

RF2. Agrupación de costos y consumos del servicio de energía eléctrica por totales y promedios.

RF3. Consumos y costos categorizados por los días de la semana(días ordinarios, días festivos y fines de semana).

RF4. Consumos y costos categorizados por las franjas del día.

RF5. Consumos y costos categorizados según los trimestres del año.

RF6. Consumos y costos categorizados según por los meses del año.

RF7. Consumos y costos categorizados según el tipo de tarifa vigente para los usuarios, estas tarifas pueden ser estáticas(STD) o dinámicas(ToU).

RF8. Consumos y costos categorizados por demografía según los grupos y/o categorías ACORN.

Requerimientos no funcionales

RNF1. Se debe permitir la combinación de filtros de diferentes tipos.

RNF2. Se debe mostrar la información completa y descripciones de los datos de cada reporte.

RNF3. Se debe mostrar los datos sobre los cuales se genera cada gráfica tabulados.

RNF4. Se debe permitir la impresión y/o exportación de gráficas y datos de cada reporte generado.

Tecnologías de desarrollo

El desarrollo del módulo de visualización de reportes se realizó bajo el modelo vista controlador(MVC). Para la vista se utilizó Bootstrap[36] que es una librería de código abierto desarrollada con HTML[37], CSS[38] y JavaScript[39] que ofrece funcionalidades para el diseño de sitios y aplicaciones web, para la representación de los datos en gráficas se uso Chart.js[40] que es una librería de código abierto basada en JavaScript que permite muchas facilidades en la visualización de gráficos estadísticos, además de usar de HTML, CSS y JavaScript para la adición de funcionalidades y diseños propios. Para el modelo y el controlador se uso PHP[41] y se implementaron las funcionalidades de consulta de la información a graficar y para funciones de conexión con la bodega de datos.

Implementación de módulo de reportes

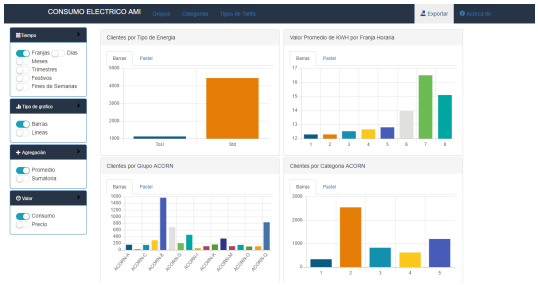
Se desarrolló una aplicación web que proporciona información sobre el consumo eléctrico residencial agrupada según factores demográficos y tarifaríos. En los factores demográficos se puede seleccionar entre las categorías o grupos ACORN y en los factores de tarifa se puede graficar según el tipo de tarifa. Para ambos tipos de agrupación se pueden aplicar los siguientes filtros de datos:

1. **Tiempo:** Permite subcategorizar la información por franjas, días, meses, trimestres, y tipos de días(festivos, fines de semana y días ordinarios).

2. **Agregación:** Se permite agregar los datos en promedio o sumatoria(total) del consumo.

3. **Tipo de Gráfico:** La información puede visualizarse en gráfico de barras, lineas, circular o de anillo.

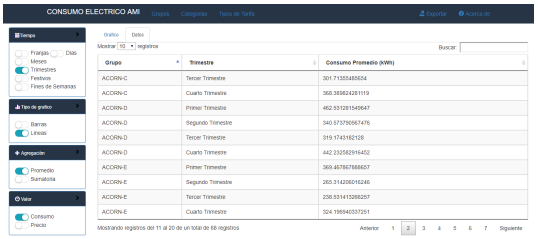
4. **Valor:** Se permite graficar consumo o costo del consumo.



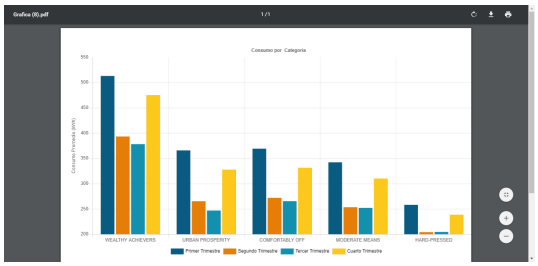
(a) Inicio.

(b) Gráfico de barras.

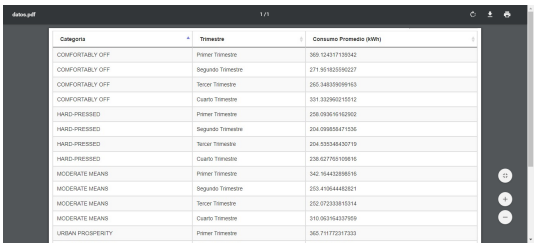
(c) Gráfico de lineas.



(d) Tabla de datos.



(e) Exportación de gráfica.



(f) Exportación de tabla de datos.

Figura 20: Interfaz del Modulo de Visualización.

Adicionalmente, se implementó la visualización tabular de los datos filtrados, exportación de gráficas y tablas de datos.

14. Pruebas

En el proceso de pruebas se debe validar el cumplimiento de cada uno de los requerimientos funcionales y no funcionales del modulo de visualización de reportes.

Plan de Pruebas

Formato de pruebas

Prueba de Software No. ##	
Requerimientos	Requerimientos a probar.
Fecha	Fecha de realización de la prueba en formato DD/MM/YYYY.
Tipo de prueba	Tipo de prueba a realizar.
Objetivo	Descripción en detalle de la funcionalidad de la prueba.
Procedimiento	Pasos a seguir para cumplir con el objetivo de la prueba.
Resultados esperados	Resultados que debería arrojar la prueba.
Resultados obtenidos	Resultados obtenidos después del proceso de ejecución de la prueba.

Tabla 10: Formato de pruebas.

Resultados de Pruebas

Prueba de Software No. 01	
Requerimientos	RF1
Fecha	01/02/2019
Tipo de prueba	Funcional
Objetivo	Validación de visualización de gráficos de datos según grupos Acorn, categorías Acorn o tipos de tarifa
Procedimiento	1. Abrir el DashBoard AMI. 2. Seleccionar una opción en cada filtro de la barra de opciones lateral. 3. Dar clic en las opciones del menú de navegación superior para la generación de los gráficos por grupos Acorn, categorías Acorn o tipos de tarifa respectivamente.
Resultados esperados	1. Se debe visualizar una etiqueta de datos en el eje X del gráfico por cada grupo Acorn, categoría Acorn o tipos de tarifa según se haya seleccionado. 2. En el titulo del gráfico se debe describir que opción del menú de navegación fue seleccionada.
Resultados obtenidos	Prueba Exitosa.

Tabla 11: Resultado de pruebas RF1.

Prueba de Software No. 02	
Requerimientos	RF2
Fecha	01/02/2019
Tipo de prueba	Funcional
Objetivo	Validación de la agrupación de costos y consumos del servicio de energía eléctrica por totales y promedios.
Procedimiento	1. Abrir el DashBoard AMI. 2. Seleccionar en la barra de opciones lateral en el filtro de agregación las opciones de Promedio o Sumatoria respectivamente. 3. Seleccionar una opción en cada filtro restante de la barra de opciones lateral. 4. Dar clic en alguna de las opciones del menú de navegación.
Resultados esperados	1. El titulo del eje Y del gráfico debe describir el tipo de agregación seleccionado. 2. Los valores graficados deben coincidir con el promedio de los datos según los filtros y opciones seleccionadas.
Resultados obtenidos	Prueba Exitosa

Tabla 12: Resultado de pruebas RF2.

Prueba de Software No. 03	
Requerimientos	RF3
Fecha	01/02/2019
Tipo de prueba	Funcional
Objetivo	Validar el funcionamiento de filtros de consumos y costos categorizados por los días de la semana(días ordinarios, días festivos y fines de semana).
Procedimiento	1. Abrir el DashBoard AMI. 2. Seleccionar en la barra de opciones lateral en el filtro de Valor las opciones de Precio o Consumo respectivamente. 3. Seleccionar en la barra de opciones lateral en el filtro Tiempo las opciones Días, Festivos y Fines de Semana respectivamente. 4. Seleccionar una opción en cada filtro restante de la barra de opciones lateral 5. Dar clic en alguna de las opciones del menú de navegación.
Resultados esperados	1. Los títulos del gráfico y del eje Y deben describir el tipo de valor seleccionado. 2. Las etiquetas del eje Y del gráfico deben coincidir con este tipo de valor. 3. En la leyenda del gráfico se debe describir los valores graficados por cada opción del filtro de tiempo seleccionado, para Días (Lunes, Martes, Miércoles, Jueves, Viernes, Sábado y Domingo), Festivos(No festivos y Festivos) y Fines de Semana(Día semana y Fin de Semana).
Resultados obtenidos	Prueba Exitosa.

Tabla 13: Resultado de pruebas RF3.

Prueba de Software No. 04	
Requerimientos	RF4
Fecha	01/02/2019
Tipo de prueba	Funcional
Objetivo	Validar el funcionamiento de filtros de consumos y costos categorizados por las franjas del día.
Procedimiento	1. Abrir el DashBoard AMI. 2. Seleccionar en la barra de opciones lateral en el filtro de Valor las opciones de Precio o Consumo respectivamente. 3. Seleccionar en la barra de opciones lateral en el filtro Tiempo la opción Franja. 4. Seleccionar una opción en cada filtro restante de la barra de opciones lateral. 5. Dar clic en alguna de las opciones del menú de navegación.
Resultados esperados	1. Los títulos del gráfico y del eje Y deben describir el tipo de valor seleccionado. 2. Las etiquetas del eje Y del gráfico deben coincidir con este tipo de valor. 3. En la leyenda del gráfico se debe describir los valores graficados de la opción Franja del filtro de tiempo seleccionado mostrando la hora inicial y final de la franja (00:00 - 02:59, 03:00 - 05:59, 06:00 - 08:59, 09:00 - 11:59, 12:00 - 14:59, 15:00 - 17:59, 18:00 - 20:59, 21:00 - 23:59).
Resultados obtenidos	Prueba exitosa.

Tabla 14: Resultado de pruebas RF4.

Prueba de Software No. 05	
Requerimientos	RF5
Fecha	01/02/2019
Tipo de prueba	Funcional
Objetivo	Validar el funcionamiento de filtros de consumos y costos categorizados según los trimestres del año.
Procedimiento	1. Abrir el DashBoard AMI. 2. Seleccionar en la barra de opciones lateral en el filtro de Valor las opciones de Precio o Consumo respectivamente. 3. Seleccionar en la barra de opciones lateral en el filtro Tiempo la opción Trimestres. 4. Seleccionar una opción en cada filtro restante de la barra de opciones lateral. 5. Dar clic en alguna de las opciones del menú de navegación.
Resultados esperados	1. Los títulos del gráfico y del eje Y deben describir el tipo de valor seleccionado. 2. Las etiquetas del eje Y del gráfico deben coincidir con este tipo de valor. 3. En la leyenda del gráfico se debe describir los valores graficados de la opción Trimestres del filtro de tiempo seleccionado(Primer Trimestre, Segundo Trimestre, Tercer Trimestre, Cuarto Trimestre).
Resultados obtenidos	Prueba exitosa.

Tabla 15: Resultado de pruebas RF5.

Prueba de Software No. 06	
Requerimientos	RF6
Fecha	01/02/2019
Tipo de prueba	Funcional
Objetivo	Validar el funcionamiento de filtros de consumos y costos categorizados según por los meses del año.
Procedimiento	1. Abrir el DashBoard AMI. 2. Seleccionar en la barra de opciones lateral en el filtro de Valor las opciones de Precio o Consumo respectivamente. 3. Seleccionar en la barra de opciones lateral en el filtro Tiempo la opción Meses. 4. Seleccionar una opción en cada filtro restante de la barra de opciones lateral. 5. Dar clic en alguna de las opciones del menú de navegación.
Resultados esperados	1. Los títulos del gráfico y del eje Y deben describir el tipo de valor seleccionado. 2. Las etiquetas del eje Y del gráfico deben coincidir con este tipo de valor. 3. En la leyenda del gráfico se debe describir los valores graficados de la opción Meses del filtro de tiempo seleccionado(Enero, Febrero, Marzo, Abril, Mayo, Junio, Julio, Agosto, Septiembre, Octubre, Noviembre y Diciembre).
Resultados obtenidos	Prueba exitosa.

Tabla 16: Resultado de pruebas RF6.

Prueba de Software No. 07	
Requerimientos	RF7
Fecha	01/02/2019
Tipo de prueba	Funcional
Objetivo	Validar el funcionamiento de filtros de consumos y costos categorizados según el tipo de tarifa vigente para los usuarios, estas tarifas pueden ser estáticas(STD) o dinámicas(ToU).
Procedimiento	1. Abrir el DashBoard AMI. 2. Seleccionar en la barra de opciones lateral en el filtro de Valor las opciones de Precio o Consumo respectivamente. 3. Seleccionar una opción en cada filtro restante de la barra de opciones lateral. 4. Dar clic en la opción Tipos de Tarifa del menú de navegación.
Resultados esperados	1. Los títulos del gráfico y del eje Y deben describir el tipo de valor seleccionado. 2. Las etiquetas del eje Y del gráfico deben coincidir con este tipo de valor. 3. El titulo del gráfico debe mostrar que se esta graficando según el tipo de tarifa. 4. En las etiquetas del eje X del gráfico se debe describir los valores de los tipos de tarifas(Std y Tou).
Resultados obtenidos	Prueba exitosa.

Tabla 17: Resultado de pruebas RF7.

Prueba de Software No. 08	
Requerimientos	RF8
Fecha	01/02/2019
Tipo de prueba	Funcional
Objetivo	Validar el funcionamiento de filtros de consumos y costos categorizados por demografía según los grupos y/o categorías ACORN.
Procedimiento	1. Abrir el DashBoard AMI. 2. Seleccionar en la barra de opciones lateral en el filtro de Valor las opciones de Precio o Consumo respectivamente. 3. Seleccionar una opción en cada filtro restante de la barra de opciones lateral. 4. Dar clic en las opciones grupos Acorn y Categorías Acorn del menú de navegación respectivamente.
Resultados esperados	1. Los títulos del gráfico y del eje Y deben describir el tipo de valor seleccionado. 2. Las etiquetas del eje Y del gráfico deben coincidir con este tipo de valor. 3. El titulo del gráfico debe mostrar que se esta graficando según la característica demográfica de Acorn correspondiente(Categorías o Grupos). 4. En las etiquetas del eje X del gráfico se debe describir los valores de las características o los grupos existentes en esta clasificación.
Resultados obtenidos	Prueba exitosa.

Tabla 18: Resultado de pruebas RF8.

Prueba de Software No. 10	
Requerimientos	RNF1
Fecha	01/02/2019
Tipo de prueba	No Funcional
Objetivo	Validar funcionamiento de filtros de diferentes tipos y sus posibles combinaciones.
Procedimiento	1. Intentar desmarcar todas las opciones en cada menú de la barra de opciones lateral. 2. Intentar marcar mas de una opción en cada menú en la barra de opciones lateral.
Resultados esperados	1. Solo se debe permitir la selección de una opción por menú de la barra de opciones lateral. 2. Al marcar una opción de un menú de la barra de opciones lateral, todas las demás opciones pertenecientes al mismo menú deben desmarcarse automáticamente.
Resultados obtenidos	Prueba exitosa.

Tabla 19: Resultado de pruebas RNF1.

Prueba de Software No. 11	
Requerimientos	RNF2
Fecha	01/02/2019
Tipo de prueba	No Funcional
Objetivo	Validar la correcta visualización de opciones de cada gráfico generado.
Procedimiento	1. Abrir el DashBoard AMI 2. Generar los diferentes tipos de gráficos y realizar la validación de los siguientes elementos en cada uno: *a*) Titulo del gráfico. *b*) Titulo del eje Y. *c*) Etiquetas del eje X y Y. *d*) Leyenda del gráfico. *e*) Barras y/o lineas de datos según sea el caso.
Resultados esperados	1. Visualización de todos los elementos del gráfico. 2. Claridad en la información representada.
Resultados obtenidos	Prueba exitosa.

Tabla 20: Resultado de pruebas RNF2.

Prueba de Software No. 12	
Requerimientos	RNF3
Fecha	01/02/2019
Tipo de prueba	No Funcional
Objetivo	Validar funcionamiento de tabulación de datos generados en cada consulta para la generación de gráficos.
Procedimiento	1. Abrir el DashBoard AMI. 2. Generar gráficos con las diferentes combinaciones de filtros de opciones de la barra lateral y del menú de navegación. 3. Dar clic en la pestaña Datos.
Resultados esperados	1. Se debe visualizar un tabla con tres columnas la primera y la segunda con los valores posibles de las opciones seleccionadas en la barra de navegación y el menú de Tiempo en la barra de opciones lateral respectivamente, y la tercera con los valores graficados por cada combinación de las anteriores columnas. 2. Se debe permitir el ordenamiento de los datos. 3. Se debe permitir el filtrado de información. 4. Se debe permitir la selección de registros a visualizar por pagina y la navegación entre las paginas existentes.
Resultados obtenidos	Prueba exitosa.

Tabla 21: Resultado de pruebas RNF3.

Prueba de Software No. 13	
Requerimientos	RNF4
Fecha	01/02/2019
Tipo de prueba	No Funcional
Objetivo	Validar funcionamiento exportación e impresión de gráficos y datos en tablas de los diferentes reportes.
Procedimiento	1. Abrir el DashBoard AMI. 2. Se seleccionan las diferentes opciones en los menús de la barra de opciones lateral. 3. Se da clic en alguna opción de la barra de navegación. 4. Se verifica la correcta visualización de la gráfica y tabla de datos. 5. Se da clic en la opción Exportar de la barra de navegación.
Resultados esperados	1. Se debe generar el archivo de exportación en formato PDF(.pdf) para descargar. 2. En la primera pagina del archivo se debe visualizar la gráfica y en las paginas restantes se debe visualizar la información tabulada.
Resultados obtenidos	Prueba exitosa.

Tabla 22: Resultado de pruebas RNF4.

Plan de Despliegue

Con el objetivo de facilitar los procesos de gestión y toma de decisiones sobre sistemas con grandes volúmenes de datos generados por sistemas AMI se implemento una aplicación web que permite la generación de informes del consumo de energía eléctrica. En esta sección se describe el objetivo de la implementación de esta solución en el entorno del usuario final. El sistema de análisis de datos de consumo debe estar disponible de forma remota para los usuarios. Además, se debe realizar el proceso de procesamiento de nuevos datos con una periodicidad de tiempo tal que permita la visualización de informacion actualizada en los reportes generados sin sobrecargar el sistema de análisis

y gestión. Es de gran importancia proveer al usuario final del sistema los mecanismos de carga de nuevos datos, con el objetivo de generar autonomía en dichos procesos. El sistema debe ser montado en servidores que soporten PHP. Los procesos de datos cargan la información proveniente de las fuentes de datos provistas en un esquema de bases de datos en PostgresSQL.

15. Conclusiones

En este trabajo se presenta el proceso de implementación de una posible solución al problema de análisis de consumos de energía eléctrica residenciales medidos con sistemas AMI. Como resultado de la implementación de tecnologías de inteligencia de negocios y la minería de datos, se obtiene una aplicación web que provee la visualización de reportes de consumos por características demográficas, temporales y tarifarías. El desarrollo de esta aplicación es de gran utilidad para la industria eléctrica, específicamente el campo de la prestación del servicio de energía eléctrica residencial, ya que provee un marco de gestión y análisis de grandes cantidades de datos resultantes de la implementación de tecnologías de medición inteligentes de consumos. Estas tecnologías han permitido que los usuarios del servicio de energía eléctrica asuman un rol de gran importancia en el estudio de estrategias de optimización de consumo y la implementación de esquemas de oferta del servicio.

Por medio de este trabajo se vislumbra una vez mas lo grandes beneficios obtenidos de la implementación de técnicas de inteligencia de negocios y minería de datos en el sector eléctrico, ya que además de facilitar la gestión de grandes volúmenes de datos, estas técnicas tienen la característica de simplificar problemas de análisis en las diversas áreas de conocimiento.Históricamente se sabe que el conocimiento provee gran variedad de ventajas competitivas a las organizaciones y es por esto que la implementación de técnicas que generen este conocimiento a partir de datos que a simple vista pueden carecer de significado es altamente valorado.

Durante la implementación de este trabajo se presentaron diferentes retos que permitieron enriquecer la experiencia como profesional en el área de sistemas de información. Algunos de estos retos fueron el aprendizaje de nuevas tecnologías de desarrollo y gestión de bases de datos. Además se permitió la puesta en práctica de conocimientos adquiridos a lo largo de la carrera en campos como el desarrollo de software, gestión de bases de datos y descubrimiento de conocimiento de base de datos.

Algunos trabajos futuros con base en el trabajo realizado pueden ser la implementación del análisis de variables ambientales en los patrones de consumo de energía eléctrica en usuarios residenciales, incluir el análisis de consumo de grandes clientes del servicio de energía eléctrica y aplicar técnicas de aprendizaje supervisado que permitan la generación de modelos descriptivos de patrones de consumo mas precisos.

Referencias

[1] Fernanda Aponte. Diseño de una base de datos. http://fernandaaponte1998.blogspot.com/2015/12/diseno-de-una-base-de-datos.html. Consultado: 2017-11-30.

[2] Jose Siguenas. Minería de datos. http://wwwmineradedatos.blogspot.com/. consultado: 2017-11-30.

[3] Hadoop architecture overview. hadoop internals. http://ercoppa.github.io/HadoopInternals/HadoopArchitectureOverview.html. Consultado: 2017-11-30.

[4] Business intelligence: Motivity solutions. https://motivitysolutions.com/business-intelligence/. Consultado: 2017-11-30.

[5] The increasing importance of security for the smart grid. http://www.elp.com/articles/powergrid_international/print/volume-16/issue-4/features/the-increasing-importance-of-security-for-the-smart-grid.html. Consultado: 2017-11-30.

[6] Replicación multimaestro alwayson para bases de datos postgresql distribuidas. https://www.2ndquadrant.com/es/recursos/postgres-bdr-2ndquadrant/. Consultado: 201-01-12.

[7] Sharding with postgresql. https://blog.dbi-services.com/sharding-with-postgresql/. Consultado: 2018-10-03.

[8] Understanding consumers and communities. https://acorn.caci.co.uk/what-is-acorn. Consultado: 2018-11-03.

[9] Hahn Tram and Hahn Tram. Enterprise information & process change management for ami and demand response. In *IEEE PES T&D 2010*, 2010.

[10] German Arce Zapata. Decreto 348 de 2017. *Ministerio de Minas y Energía*, page 6.

[11] M Weiser, R Gold, and J S Brown. The origins of ubiquitous computing research at PARC in the late 1980s. *IBM Syst. J.*, 38(4):693–696, 1999.

[12] J.H. Orallo, M.J.R. Quintana, and C.F. Ramírez. *Introducción a la minería de datos*. Fuera de colección Out of series. Pearson Educación, 2004.

[13] R Adapa. Expert system applications in power system planning and operations. *IEEE Power Engineering Review*, 14(2):12, 1994.

[14] Xu Tao, Xu Tao, He Renmu, Wang Peng, and Xu Dongjie. Applications of data mining technique for power system transient stability prediction. In *2004 IEEE International Conference on Electric Utility Deregulation, Restructuring and Power Technologies. Proceedings.*

[15] Fengyu Wang Yanliu Cheng Houlei Lv, Yingxin Xie. Research on intelligent power consumption business intelligence system based on cloud computing. In *Intelligent Information Hiding and Multimedia Signal Processing, 2013 Ninth International Conference on*. IEEE, December 2012.

[16] N. Jacome-Grajales, G. Escobedo-Briones, J. Roblero and G. Arroyo-Figueroa. Application of business intelligence to the power system process security. November 2013.

[17] Qiong Ren and Jun Tao. KPI corporate management and business intelligence analysis on the application of electric power enterprises. In *2012 International Conference on Industrial Control and Electronics Engineering*, 2012.

[18] Seon Yeong Han, Jaegoo No, Yongjae Joo, and Jin-Ho Shin. Conditional abnormality detection based on AMI data mining. *IET Gener. Transm. Distrib.*, 10(12):3010–3016, 2016.

[19] Chen Rui, Hou Yibin, Huang Zhangqin, and He Jian. Data management model for ambient intelligence: AmI-Data. In *2009 WRI International Conference on Communications and Mobile Computing*, 2009.

[20] Wanrong Qiu, Feng Zhai, Zhejing Bao, Baofeng Li, Qiang Yang, and Yongfeng Cao. Clustering approach and characteristic indices for load profiles of customers using data from AMI. In *2016 China International Conference on Electricity Distribution (CICED)*, 2016.

[21] Elzbieta Malinowski and Esteban Zimányi. *Advanced Data Warehouse Design: From Conventional to Spatial and Temporal Applications*. Springer Science & Business Media, January 2008.

[22] E F Codd. A relational model of data for large shared data banks. 1970. *MD Comput.*, 15(3):162–166, May 1998.

[23] Iggy Fernandez. *Beginning Oracle Database 11g Administration*. 2009.

[24] Ma Victoria Nevado Cabello. *Introduccion a Las Bases de Datos Relacionales*. Editorial Visión Libros.

[25] Ming-Syan Chen, Ming-Syan Chen, Jiawei Han, and P S Yu. Data mining: an overview from a database perspective. *IEEE Trans. Knowl. Data Eng.*, 8(6):866–883, 1996.

[26] Osmar Zaiane. Chapter 1: Introduction to data mining. `https://webdocs.cs.ualberta.ca/~zaiane/courses/cmput690/notes/Chapter1/index.html#s3`, September 1999. Consultado: 2017-11-29.

[27] Patrick C K Hung. *Big Data Applications and Use Cases*. Springer, May 2016.

[28] Apache hadoop. http://hadoop.apache.org/#What+Is+Apache+Hadoop%3F. Consultado: 2017-11-27.

[29] Josep Curto Díaz. *Introducción al Business Intelligence*. Editorial UOC, May 2012.

[30] Superintendencia de Industria y Comercio Centro de Información Tecnológica Y Apoyo A La Gestión de La Propiedad Industrial (cigepi). Medición y Gestión Inteligente de ConsumoEléctrico. page 96, 2016.

[31] J.V.T.S.A.N. MIGUEL. *UF1470 - Administración y monitorización de los SGBD*. Ediciones Paraninfo, S.A, 2016.

[32] Dongkuan Xu and Yingjie Tian. A comprehensive survey of clustering algorithms. *Annals of Data Science*, 2(2):165–193, Jun 2015.

[33] Irene García Cambronero, Cristina; Gómez Moreno. Algoritmos de aprendizaje: Knn and kmeans. *http://blogs.ujaen.es/barranco/wp-content/uploads/2012/02/Algoritmos-de-aprendizaje-knn-y-kmeans.pdf*, 2012.

[34] The University Waikato. Weka 3: Data mining software in java. https://www.cs.waikato.ac.nz/ml/weka/. Consultado: 2018-11-23.

[35] M A Syakur, B K Khotimah, E M S Rochman, and B D Satoto. Integration k-means clustering method and elbow method for identification of the best customer profile cluster. *IOP Conference Series: Materials Science and Engineering*, 336:012017, apr 2018.

[36] Bootstrap. https://getbootstrap.com/. Consultado: 2019-14-13.

[37] Html5 introduction. https://www.w3schools.com/html/html5_intro.asp. Consultado: 2019-04-13.

[38] Css tutorial. https://www.w3schools.com/css/default.asp. Consultado: 2019-04-13.

[39] Javascript tutorial. https://www.w3schools.com/js/. Consultado: 2019-04-13.

[40] Chart.js open source html5 charts for your website. https://www.chartjs.org/. Consultado: 2019-04-13.

[41] Programación en php. https://es.wikibooks.org/wiki/Programacion_en_PHP. Consultado: 2019-04-13.

[42] Sharding your data with postgresql 11. https://pgdash.io/blog/postgres-11-sharding.html. Consultado: 2018-11-13.

Anexos

Implementación de replicación con PostgresBDR

Conceptos básicos

Instalación de PostgreSQL

Se crea un archivo de fuentes donde incluimos la fuente de PostgreSQL para nuestra distribución de Linux.Agregamos el archivo PGDB APT Source y editamos el archivo en nuestro editor elegido(En este caso nano), y agregamos la linea "deb http://apt.postgresql.org/pub/ xenial-pgdg main.[es]pecificando la distribucion de linux (en este caso xenial).

```
1 $ sudo touch /etc/apt/sources.list.d/pgdg.list
2 $ sudo nano /etc/apt/sources.list.d/pgdg.list
```

Agregamos el origen del paquete con el comando.

```
1 $sudo sh -c 'echo "deb http://apt.postgresql.org/pub/repos/apt
    / trusty-pgdg main" > /etc/apt/sources.list.d/pgdg.list '
```

Agregamos la clave del repositorio de PostgreSQL e instalamos PostgreSQL version 9.4.

```
1 $ sudo apt-get install wget ca-certificates
2 $ wget --quiet -O - https://www.postgresql.org/media/keys/
    ACCC4CF8.asc | sudo apt-key add -
3 $ sudo apt-get update
4 $ sudo apt-get upgrade
5 $ sudo apt-get install postgresql-9.4 pgadmin3
```

Instalación de BDR

Agregamos el archivo PGDB APT Source y lo editamos agregando la fuente de Postgres-BDR con la linea "deb [arch=amd64] http://packages.2ndquadrant.com/bdr/apt/ xenial-2ndquadrant main".

```
1 $ sudo touch /etc/apt/sources.list.d/2ndquadrant.list
2 $ sudo nano /etc/apt/sources.list.d/2ndquadrant.list
```

Ahora instalamos BDR y exportamos archivos binarios que nos permitir abreviar la ejecución de funciones de PostgreSQL.

```
1 $ wget --quiet -O - http://packages.2ndquadrant.com/bdr/apt/
    AA7A6805.asc | sudo apt-key add -
2 $ sudo apt-get update
```

```
3 $ sudo apt-get install postgresql-bdr-9.4-bdr-plugin
4 $ export PATH=/usr/lib/postgresql/9.4/bin:$PATH
```

Listing 1: Instalación de prueba de BDR

```
1 $ curl -s "https://raw.githubusercontent.com/2ndQuadrant/bdr/
    bdr-plugin/REL1_0_STABLE/scripts/bdr_quickstart.sh" | bash
2 $ export PATH=$HOME/2ndquadrant_bdr/bdr/bin:$PATH
```

Creación de Nodos

Se usa el comando initdb para la creación de nodos de PostgresSQL que son grupos de bases de datos administradas por una sola instancia de servidor. En el initbd debemos especificar el directorio, nombre del nodo, método de autenticación y nombre del usuario.

```
1 $ initdb -D DIRECTORIO/NOMBRE_NODO -A METODO-AUT -U USUARIO
```

Donde se especifican:
-D DIRECTORIO: Directorio de almacenamiento del nodo(opción obligatoria).
-A METODO-AUT: Método de autenticación para usuarios locales.
-U USUARIO: Selecciona el superusuario de la base de datos.

Edición de Archivos de Configuración de Nodos

Configuración de postgresql.conf El archivo postgresql.conf contiene parámetros de configuración del comportamiento de PostgreSQL en nuestro nodo. Los parametros que modificaremos son los siguientes:
shared preload libraries: Permite precargar bibliotecas compartidas en el servidor.
wal level: Determina como es escrita la informacion en el WAL.las opciones de configuracion son replica, minimal y logical.
track commit timestamp: Define si se guarda el tiempo de confirmacion de las transacciones.
max connections: Numero máximo de clientes conectados a la vez en nuestra base de datos
max wal senders: Cantidad máxima de conexiones simultaneas desde servidores
max replication slots : Cantidad máxima de ranuras de replicaciones
max worker processes: Cantidad maxima de procesos en segundo plano admitidos por el sistema

NOTA. *No solo es necesario modificar las opciones anteriormente mencionadas, sino también comentar algunas que no son necesarias como log error verbosity, log min*

messages, log line prefix, log line prefix, bdr.default apply delay y bdr.log conflicts to table.

Listing 2: Configuracion de postgresql.conf

```
1   shared_preload_libraries = 'bdr'
2   wal_level = 'logical'
3   track_commit_timestamp = on
4   max_connections = 100
5   max_wal_senders = 10
6   max_replication_slots = 10
7   # Make sure there are enough background worker slots for
        BDR to run
8   max_worker_processes = 10
9
10  # These aren't required, but are useful for diagnosing
        problems
11  #log_error_verbosity = verbose
12  #log_min_messages = debug1
13  #log_line_prefix = 'd=%d p=%p a=%a%q '
14
15  # Useful options for playing with conflicts
16  #bdr.default_apply_delay=2000   # milliseconds
17  #bdr.log_conflicts_to_table=on
```

Configuración de pg-hba.conf Este archivo contiene información de accesos para los diferentes usuarios que pueden conectarse a nuestro nodo; los registros se almacenan uno en cada linea y contienen el tipo de conexión, rango de dirección IP del cliente, nombre del usuario, nombre de la base de datos y el tipo de autenticacion.

Listing 3: Archivo pg-hba.conf

```
1  local   replication   postgres                    trust
2  host    replication   postgres    127.0.0.1/32    trust
3  host    replication   postgres    ::1/128         trust
```

Levantamiento de nodos

Para levantar nodos se usa el comando pg-ctl que es usado para inicializar, iniciar, detener o controlar servidores de PostgreSQL asi:

```
1  $pg_ctl -l ARCHIVO -D DIR-NODO -o "-p DIR" -w start
```

-l ARCHIVO: Ubicación de archivo de salida log.
-D DIR-NODO: Ubicación de los archivos de configuración.

-o : Opciones pasadas directamente a postgres, deben estar encerradas en comillas dobles.
-p DIR: Ubicación del ejecutable de postgres.
-w : Espera a que inicie o termine para continuar.

Creación de bases de datos

Ahora debemos crear la base de datos a replicar en todos los nodos usando el comando createdb.

```
1 $ createdb -p PUERTO -U USUARIO NOMBREBD
```

-p PUERTO : Especifica el puerto TCP o la extensión de archivo de socket de dominio local de Unix en el que el servidor está escuchando conexiones.
-u USUARIO : Nombre del superusuario de la base de datos.
NOMBREBD : Nombre que se asignara a la nueva base de datos.

Para acceder a las bases de datos creadas en cada nodo usamos el comando psql y le especificamos el puerto, el usuario y el nombre de la base de datos, asi:

```
1 $psql -p PUERTO -U USUARIO BASE-DE-DATOS
```

Habilitación de PostgresBDR

Para habilitar BDR instalaremos dos extensiones en cada uno de nuestros nodos por medio del comando CREATE EXTENSION que tiene el siguiente formato:

```
1 CREATE EXTENSION [ IF NOT EXISTS ] extension_name
2     [ WITH ] [ SCHEMA schema_name ]
3              [ VERSION version ]
4              [ FROM old_version ]
```

Las extensiones que instalaremos son btree-gist que proporciona clases de operador de índice GiST y bdr que nos permite realizar la replicacion.

```
1 CREATE EXTENSION btree_gist;
2 CREATE EXTENSION bdr;
```

Para realizar el enlace entre nodos, primero debemos crear el grupo en el primer nodo en el que especificamos nombre local para el nodo que debe ser exclusivo en el grupo y dsn externo del nodo que se usara para que otros nodos se conecten a este.Y nos aseguramos que el nodo este listo para replicarse.

```
1 SELECT bdr.bdr_group_create(
2     local_node_name := 'NOMBRE1',
3     node_external_dsn := 'port=PUERTO1 dbname=DB host=HOST1'
4 );
```

```
5 SELECT bdr.bdr_node_join_wait_for_ready();
```

Finalmente, enlazamos los demás nodos al grupo creado especificando nombre local, dsn externo y dsn al que se unirá nuestro nodo.

```
1 SELECT bdr.bdr_group_join(
2         local_node_name := 'NOMBRE2',
3         node_external_dsn := 'port=PUERTO2 dbname=DB host=HOST2
              ',
4         join_using_dsn := 'port=PUERTO1 dbname=DB host=HOST1'
5 );
6 SELECT bdr.bdr_node_join_wait_for_ready();
```

Prueba de sistemas PostgresBDR

La forma de probar el funcionamiento del sistema BDR es hacer conexión en alguno de los nodos, escribir o modificar información en el y verificar si la información fue actualizada en los demás nodos. Para establecer conexión con un nodo es necesario especificar en el acceso a PostgreSQL el puerto, usuario y la base de datos a la que haremos conexión en el nodo.

```
1     $psql -p PUERTO -u USUARIO BASEDEDATOS
```

Montaje de sistema de replicación

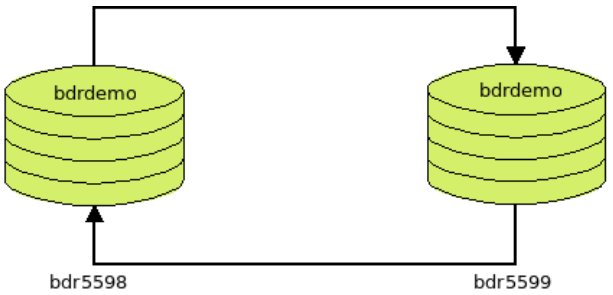


Figura 21: Sistema BDR con 2 nodos maestros

Creación de Nodos

```
1 $mkdir -p $HOME/2ndquadrant_bdr
2 $initdb -D $HOME/2ndquadrant_bdr/bdr5598 -A trust -U postgres
3 $initdb -D $HOME/2ndquadrant_bdr/bdr5599 -A trust -U postgres
```

Edición de Archivos de Configuración de Nodos

Editar en ambos nodos los siguientes archivos.

Listing 4: Archivo postgresql.conf

```
shared_preload_libraries = 'bdr'
wal_level = 'logical'
track_commit_timestamp = on
max_connections = 100
max_wal_senders = 10
max_replication_slots = 10
# Make sure there are enough background worker slots for
   BDR to run
max_worker_processes = 10

# These aren't required, but are useful for diagnosing
   problems
#log_error_verbosity = verbose
#log_min_messages = debug1
#log_line_prefix = 'd=%d p=%p a=%a %q '

# Useful options for playing with conflicts
#bdr.default_apply_delay=2000   # milliseconds
#bdr.log_conflicts_to_table=on
```

.

Listing 5: Archivo pg-hba.conf

```
local   replication   postgres                    trust
host    replication   postgres    127.0.0.1/32    trust
host    replication   postgres    ::1/128         trust
```

Levantar Nodos

```
$pg_ctl -l $HOME/2ndquadrant_bdr/bdr5598.log -D $HOME/2
   ndquadrant_bdr/bdr5598 -o "-p 5598" -w start
$pg_ctl -l $HOME/2ndquadrant_bdr/bdr5599.log -D $HOME/2
   ndquadrant_bdr/bdr5599 -o "-p 5599" -w start
```

Creación de Base de Datos de Replicación

```
$createdb -p 5598 -U postgres bdrdemo
$createdb -p 5599 -U postgres bdrdemo
```

Habilitar BDR

En el primer nodo

```
$psql -p 5598 -U postgres bdrdemo
     CREATE EXTENSION btree_gist;
     CREATE EXTENSION bdr;
```

```
SELECT bdr.bdr_group_create(
    local_node_name := 'node1',
    node_external_dsn := 'port=5598 dbname=bdrdemo host=
       localhost'
);
```

```
SELECT bdr.bdr_node_join_wait_for_ready();
```

En el segundo nodo

```
$psql -p 5599 -U postgres bdrdemo
     CREATE EXTENSION btree_gist;
     CREATE EXTENSION bdr;
```

```
SELECT bdr.bdr_group_join(
    local_node_name := 'node2',
    node_external_dsn := 'port=5599 dbname=bdrdemo host=
       localhost',
    join_using_dsn := 'port=5598 dbname=bdrdemo host=
       localhost'
);
```

```
SELECT bdr.bdr_node_join_wait_for_ready();
```

Prueba del sistemas BDR

Accedemos a la base de datos del primer nodo, creamos una tabla, adicionamos información de nuestra preferencia y consultamos si quedo correctamente almacenada en nuestra base de datos.

```
$psql -p 5598 -U postgres bdrdemo

    CREATE TABLE t1bdr (c1 INT, PRIMARY KEY (c1));
    INSERT INTO t1bdr VALUES (1);
    INSERT INTO t1bdr VALUES (2);

    SELECT * FROM t1bdr;
```

Despues verificamos si la informacion fue replicada con exito accediente a nuestro segundo nodo y ejecutando una consulta. Eliminamos informacion, verificamos si fue eliminada.

```
$psql -p 5599 -U postgres bdrdemo
    SELECT * FROM t1bdr;
    DELETE FROM t1bdr WHERE c1 = 2;
    SELECT * FROM t1bdr;
```

Finalmente, verificamos si la informacion se elimino tambien el el primer nodo.

```
$psql -p 5598 -U postgres bdrdemo
    SELECT * FROM t1bdr;
```

Implementación de fragmentación en PostgreSQL

[42]

Conceptos básicos

Se detalla brevemente los comandos necesarios para el montaje de un sistema fragmentado.

Creación de Contenedores de Datos Remotos

Para preparar el acceso remoto a los nodos del sistema instalaremos las extensión de postgres_fdw con el siguiente comando:

```
CREATE EXTENSION postgres_fdw;
```

Despues de disponer de la extensión postgres_fdw instalada, procedemos a crear la conexión con el o los servidores remotos; para esto debemos definir un nombre para el servidor remoto, dirección IP o nombre del host remoto y el nombre de la base de datos remota.

```
CREATE SERVER [name_server]
FOREIGN DATA WRAPPER postgres_fdw
  OPTIONS (host '[host]', dbname '[dbname]');
```

Ahora, realizamos la asignación del usuario que se usara para acceder a nuestros servidores remotos, esto permite usar un alias sobre un usuario local para los accesos remotos.

```
CREATE USER MAPPING FOR [user_local]
SERVER [name_server]
  OPTIONS (user '[alias_user]');
```

Implementación de Fragmentación

Con la implementación de la infraestructura de servidores remotos configurada, es muy sencillo implementar un esquema de fragmentación sobre este. Primero, creamos el esquema de la tabla a fragmentar en el servidor principal:

```
1  CREATE TABLE [name_table] (...)
2    PARTITION BY RANGE ([name_field]);
```

Se deben crear los fragmentos en los diferentes servidores remotos configurados en el esquema.

```
1  CREATE FOREIGN TABLE [name_shard]
2    PARTITION OF [name_table]
3    FOR VALUES FROM ([min]) TO ([max])
4    SERVER [name_server];
```

Con este proceso finalizado se dispone de un esquema de bases de datos fragmentado que puede ser gestionado a través de los diversos tipos de operaciones transaccionales ejecutadas sobre el servidor local o principal.

Diseño dimensional

Detalle de Dimensiones

Atributo	**Descripción**	**Valores de Muestra**
dia_id	Identificador numérico del registro.	1,2,3...
fecha	Cadena de texto con la representación de la fecha en formato yyyy-MM-dd .	2013-08-30
dia_semana	Numero del día de la semana.	1,2,...,6,7.
dia_mes	Numero del día del mes.	1,2,...,30,31.
mes	Numero del mes del año.	1,2,...,11,12.
trimestre	Trimestre del año.	1, 2,3 ó 4.
año	Numero del año.	2013
fin_semana	Numero que identifica si el día es fin de semana	0[No] ó 1[Si]
festivo	Numero que identifica si el día es festivo	0[No] ó 1[Si]
nombre_festivo	Nombre del festivo en caso de que el día sea festivo.	Spring bank holiday

Tabla 23: Detalle dimensión Día

Nombre	Descripción	Valores de Muestra
mes_id	Identificador numérico del mes.	1,2,3...
mes_anio	Numero del mes del año.	1,2,...,11,12.
trimestre	Trimestre del año.	1, 2,3 ó 4.
anio	Numero del año.	2013

Tabla 24: Detalle dimensión Mes

Nombre	Descripción	Valores de Muestra
franja_id	Identificador numérico de la franja.	1,2,3...
hora_inicio	Hora de inicio de la franja en formato hh:mm:ss	06:00:00
hora_fin	Hora de finalización de la franjas en formato hh:mm:ss	09:00:59

Tabla 25: Detalle dimensión Franja

Nombre	Descripción	Valores de Muestra
cliente_id	Identificador numérico del cliente	1,2,3...
lclid	Numero de medidor de energía eléctrica.	MAC000004
tipo_tarifa	Tipo de tarifa de consumó sujeta al cliente.	Std o ToU

Tabla 26: Detalle dimensión Cliente

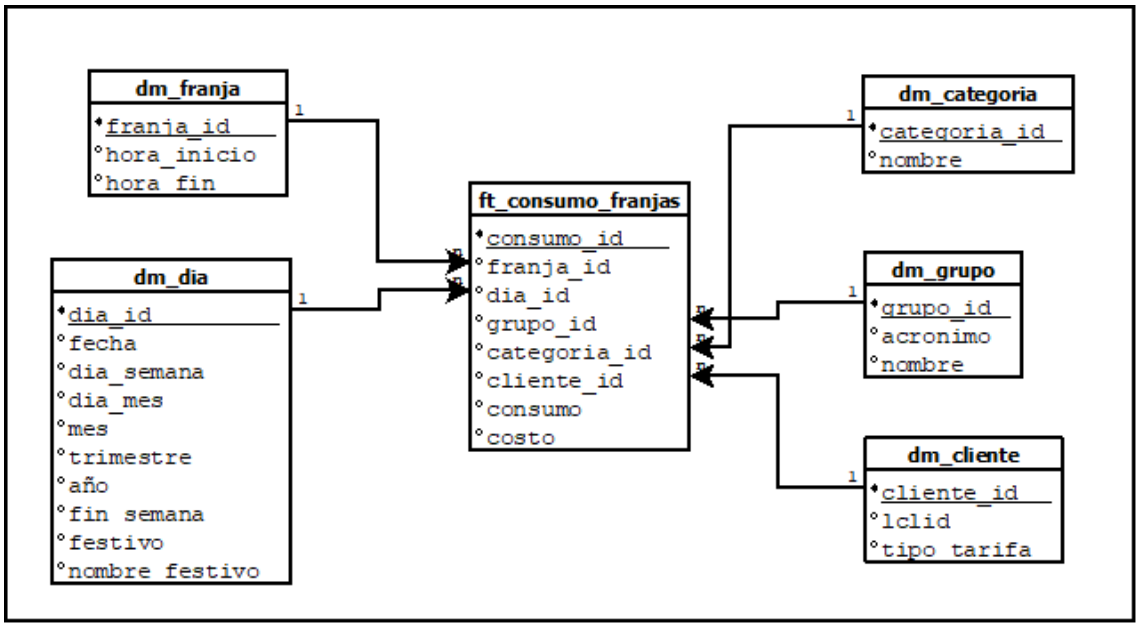
Nombre	Descripción	Valores de Muestra
grupo_id	Identificador numérico del grupo	1,2,3...
acrónimo	Nombre de la empresa que realizo la clasificación demográfica unido a la letra representativa del grupo.	ACORN-A
nombre	Nombre completo del grupo.	Wealthy Executives

Tabla 27: Detalle dimensión Grupo

Nombre	Descripción	Valores de Muestra
categoria_id	Identificador numérico de la categoría	1,2,3...
nombre	Nombre de la categoría	WEALTHY ACHIEVERS

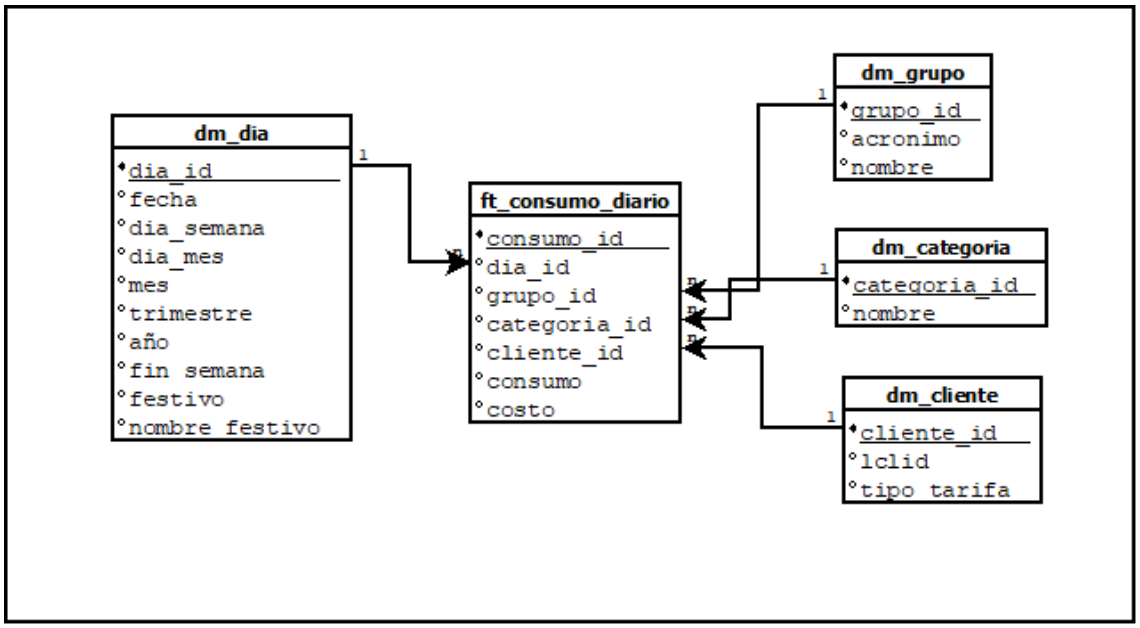
Tabla 28: Detalle dimensión Categoría

Detalle de datamarts



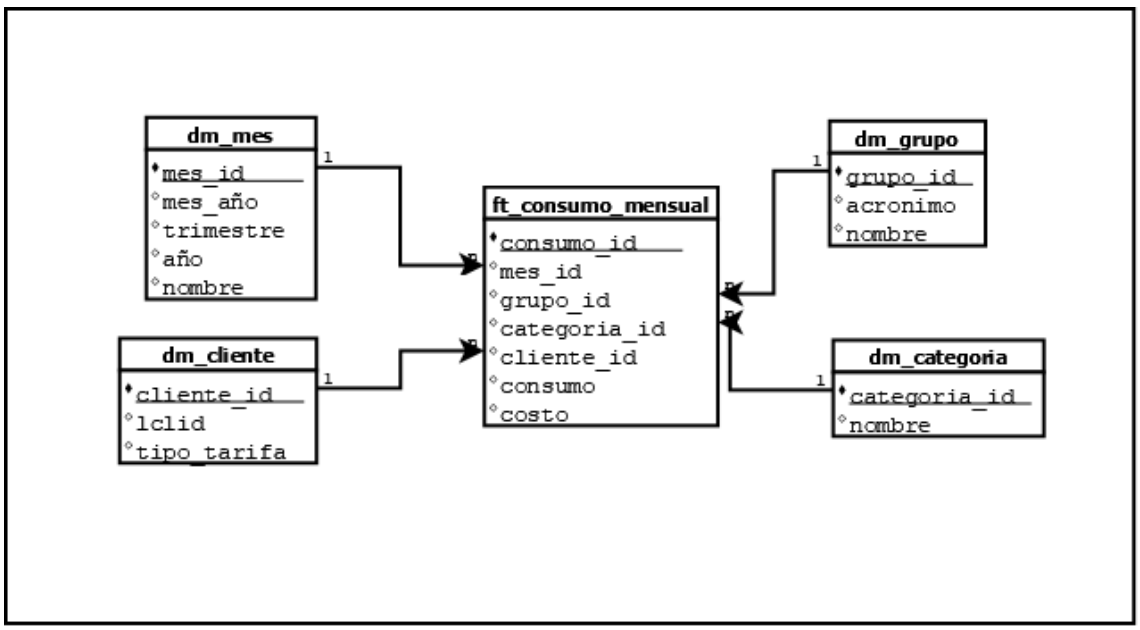
Nombre	Descripción	Regla Agregación
consumo	Consumo del usuario en la franja medido en kw/h.	Sumatorio
costo	Costo a pagar por el consumo de la franjas medido en peniques.	Sumatorio

Tabla 29: Detalle consumo franjas



Nombre	Descripción	Regla Agregación
consumo	Consumo del usuario en el día medido en kw/h.	Sumatorio
costo	Costo a pagar por el consumo del día medido en peniques.	Sumatorio

Tabla 30: Detalle consumo diario



Nombre	Descripción	Regla Agregación
consumo	Consumo del usuario en el mes medido en kw/h.	Sumatorio
costo	Costo a pagar por el consumo del mes medido en peniques.	Sumatorio

Tabla 31: Detalle consumo mensual

Diseño lógico

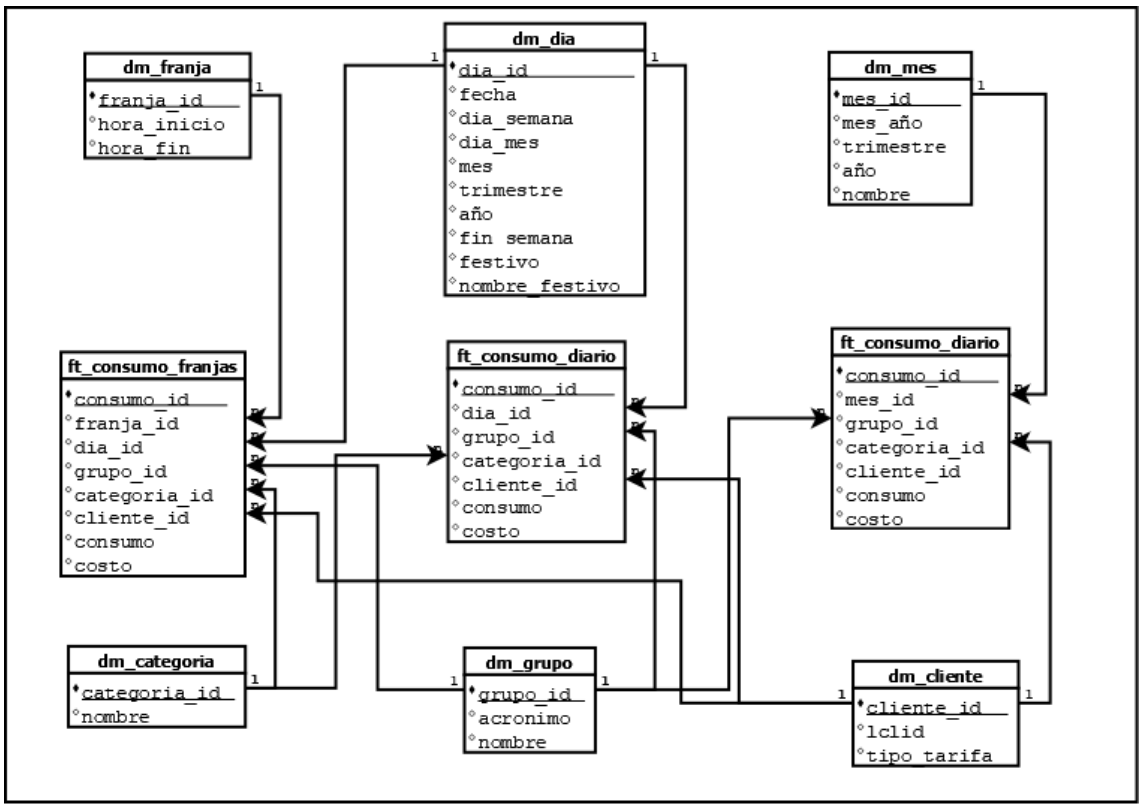


Figura 22: Diseño lógico.

Diseño físico

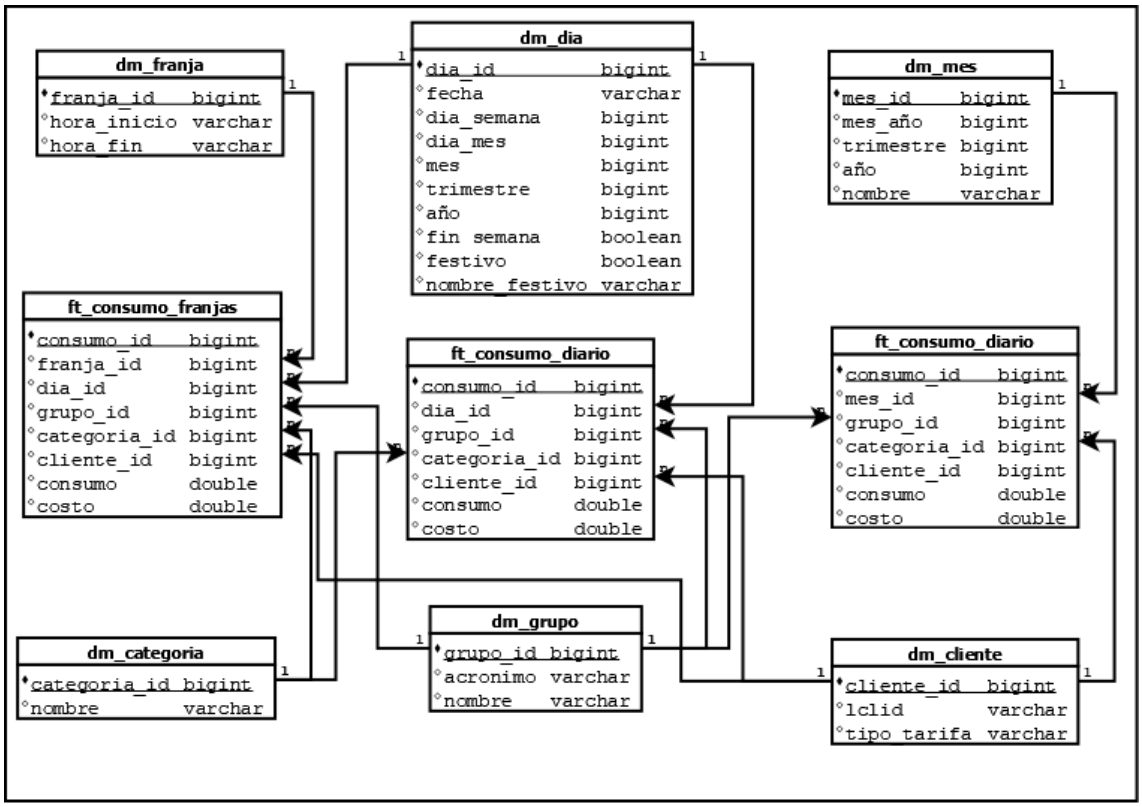


Figura 23: Diseño físico.